A NONPARAMETRIC-BASED APPROACH ON THE PROPAGATION OF IMPRECISE PROBABILITIES DUE TO SMALL DATASETS

A Thesis
Presented to
The Academic Faculty

By

Zhenyu Gao

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Aerospace Engineering

Georgia Institute of Technology

May 2018

Copyright © 2018 by Zhenyu Gao

A NONPARAMETRIC-BASED APPROACH ON THE PROPAGATION OF IMPRECISE PROBABILITIES DUE TO SMALL DATASETS

Approved by:

Dr. Dimitri N. Mavris, Advisor
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Dongwook Lim
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Katherine G. Schwartz
School of Aerospace Engineering
Georgia Institute of Technology

Date Approved: April 12, 2018

"Wind extinguishes a candle and energizes fire."

- Nassim Nicholas Taleb, Antifragile: Things That Gain from Disorder

To my parents.

ACKNOWLEDGEMENTS

Looking back to the past two years that I spent at Georgia Tech, I learned a lot from advanced level courses and challenging projects. I am glad that Georgia Tech is the place that motivates me to work harder and think deep. I want to first thank my advisor, Prof. Dimitri Mavris, for giving me the opportunity to become part of the Aerospace Systems Design Lab (ASDL). Experience at ASDL has, to a large extent, widened my vision in the aerospace industry, and made me realize the values of interdisciplinary research. Dr. Mavris' supports and encouragements have driven me to keep pursuing higher targets.

Also, I am very fortunate to have worked with Dr. Simon Briceno for all my major projects at ASDL. Dr. Briceno showed me how to get things done in a professional way, and trained me to become a better team player. I would also like to thank other ASDL research engineers whom I have worked with: Dr. Imon Chakraborty, Dr. Dongwook Lim, and Dr. Katherine Schwartz. Dr. Lim and Dr. Schwartz serve as my thesis committee members and have provided great technical guidances.

To my fellow ASDL colleagues: Yu Cai, Jiacheng Xie, Po-Nien Lin, Mingxuan Shi, Hallie Ford, Alex Lin and others, thank you so much for all the help, both in study and daily life. Especially, I want to thank Ruth Pimentel for coordinating my graduation and other applications.

To my close friends in Atlanta and elsewhere in the world: thank you for your continuous support and understanding. Without you guys, my journey would not have been so enjoyable. And thank my family, for everything.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

The quantification of uncertainty is typically done by using precise probabilities, which requires a very high level of precision and consistency of information for the uncertain sources. However, in reality, due to reasons like costs and loss of information, only small datasets or incomplete information may be obtained such that the underlying knowledge of the uncertain sources can not be adequately represented using a typical probability measure. The problem of uncertain probability distributions is also referred as imprecise probability. A common flaw in engineering community is the arbitrary or unjustified use of distributions. Assigning a probability distribution to an uncertain source based on limited knowledge and judgments has always been difficult, and incorrect assumptions of distribution types can lead to inaccuracies in the uncertainty quantification process. The incorrect assumptions, and the use of some "lack of knowledge distributions", such as Triangular distribution and Students t-distribution, may introduce unwarranted information, making uncertainty in outputs even more difficult to be quantified. Methods in imprecise probability provide greater flexibility in accommodating distributions for uncertain sources in uncertainty quantification, as they base inferences on weaker assumptions than the precise probabilistic methods, and does not require people to represent their judgments.

Latest progress in academia proposed a complete parametric-based approach for the propagation of uncertainty created by lack of sufficient statistical data. The parametric-based methods utilize multimodal inference, Bayesian inference, importance sampling and other statistical elements, and are great contributions to this area. Nevertheless, there are still some notable limitations and constraints on the parametric-based methods, such as loss of optimality, selection of candidate probability models, and constraint of well-established parametric models. This thesis work aims to formulate a nonparametric-based approach to propagate uncertainty of impre-

cise probabilities due to small datasets, while solving the constraints and limitations mentioned above.

The first part of this work is an application of nonparametric statistical methods, such as Kernel Density Estimation (KDE) and Bootstrap, to estimate the probability density function of a random variable, and provide sampling densities for uncertainty propagation. Two types of sampling densities are generated as results of this part: an optimal sampling density, and a maximum variance density. The optimal sampling density represents the best estimation for the true density based on small datasets, without overfitting and underfitting. Bandwidth selection of the optimal sampling density involves the use of different categories of methods, including rules of thumb and cross-validation. The maximum variance density provides another more conservative sampling option, which represents risk and uncertainty that is inherent in small datasets. By propagating the maximum variance densities, the objective is to simulate more potential extreme values in the output distributions and help researchers make safer decisions.

The second part extends the first part, to capture the dependencies among variables and generate multi-dimensional nonparametric density estimations. After a process to identify the correlation (strong or weak) between the variables based on small datasets, Copulas and the Sklar's Theorem are used to link the marginal nonparametric densities and create joint densities. By propagating the joint densities for dependent variables, researchers can prevent uncertainty in the outputs from being underestimated or overestimated. The effectiveness of both the 1-D and 2-D nonparametric density estimation methods are tested by selected test cases with different statistical characteristics. A complete uncertainty propagation test through a complex systems model is also conducted.

Finally, the nonparametric-based methods developed in this thesis are applied to a challenging problem in aviation environmental impact analysis, where the difficulty of

collecting complete information is a source of uncertainty. Joint distributions of some crucial input variables of the analysis are first estimated from small datasets under certain constraints, such as time, space, and data sources. In the end, an uncertainty propagation test is conducted, to estimate the distributions of fuel burn and emissions at Atlanta airport throughout a year by only using datasets from certain days. These applications show some limitations of the methods developed, yet also demonstrate the capability and ultimate target of this work: if small datasets are all that we have for uncertainty sources, this approach tries the best to still propagate uncertainty with risk consideration.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Motivation - The Reality of Small Datasets

1.1.1 Uncertainty and Uncertainty Propagation

Uncertainty is a part of our life, and it is also a part of engineering design. As a French Enlightenment writer Voltaire once said, *"Uncertainty is an uncomfortable position, but certainty is an absurd one"*. In everyday life, the existence of uncertainty may bring us difficulties in making decisions on areas such as asset management, work choice, travel destination, etc. Uncertainty is also a source of anxiety for some people, because things like the black swan events, even deemed improbable by most of the people, may happen and cause massive consequences. Two examples of the black swan events include Brexit (the U.K. departing from the European Union) in 2016, and the financial crisis in 2008. The field of uncertainty quantification (UQ) has been developed in response to the need of making better decisions under uncertainty.

People have now realized the significance of uncertainty quantification and analysis in engineering related fields, for achieving optimal decisions and risk mitigation. In a complex system that has numerous inputs and outputs, uncertainty in the outputs can not be neglected with the presence of variability in input variables. An uncertainty propagation process for a typical complex system can be used to predict the statistical properties of a non-deterministic output.

1.1.2 MCS-based Uncertainty Propagation Process

Monte Carlo Simulation (MCS) is a powerful, flexible and direct method to learn about a system by simulating it with random sampling [1]. It is named after the city

of Monte Carlo in Monaco, since the simulation process involves generating chance variables and exhibits random behaviors [2]. It is widely used in both engineering and non-engineering fields because it can deal with a large number of random variables, various distribution types, and highly nonlinear engineering models. It is often the simplest way to solve a problem, and sometimes the only feasible way [1].

Monte Carlo methods use random numbers as a tool to compute something that is not random [3]. For example, let X be a random variable with expected value $A = \mathrm{E}[X]$, we can generate n independent random variables X_1, ..., X_n to make the approximation $A \approx \hat{A}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. The strong law of large number states that $\hat{A}_n \to A$ as $n \to \infty$. In the end, the target number A is not random [3]. In the simulation process, random sampling is performed and a large number of experiments is conducted to observe statistical characteristics of outputs [2]. Then we estimate the characteristics of the output variables through analyzing the simulations results.

A basic Monte Carlo simulation process consists of three major steps (sampling, evaluating, analyzing), as shown in Figure 1.1. The Monte Carlo simulation is commonly used in the uncertainty propagation process, which is the process of mathematically mapping sources of uncertainty, from wherever they originate, to the uncertainties in the simulation results [4]. This Monte Carlo simulation-based (MCS-based) uncertainty propagation method is selected and used as the only uncertainty propagation method throughout this thesis.

Figure 1.2 illustrates the MCS-based uncertainty propagation process. This sampling based approach contains three main steps: sample the inputs (after uncertainty characterization), run the model until enough results are collected (uncertainty propagation), and gather the results (for uncertainty analysis). To further reduce the complexity and computational effort during this process, design of experiments (DoE) and surrogate modeling can be used to simplify the sophisticated original model.

Surrogate models are computationally cheaper models designed to approximate

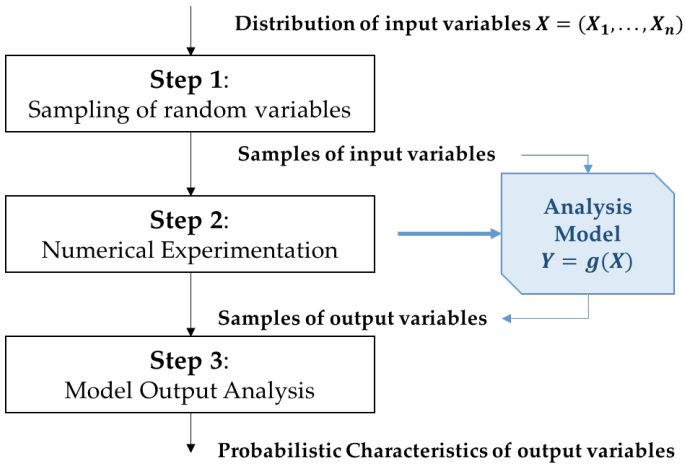


Figure 1.1: The process of a basic Monte Carlo simulation process [2]

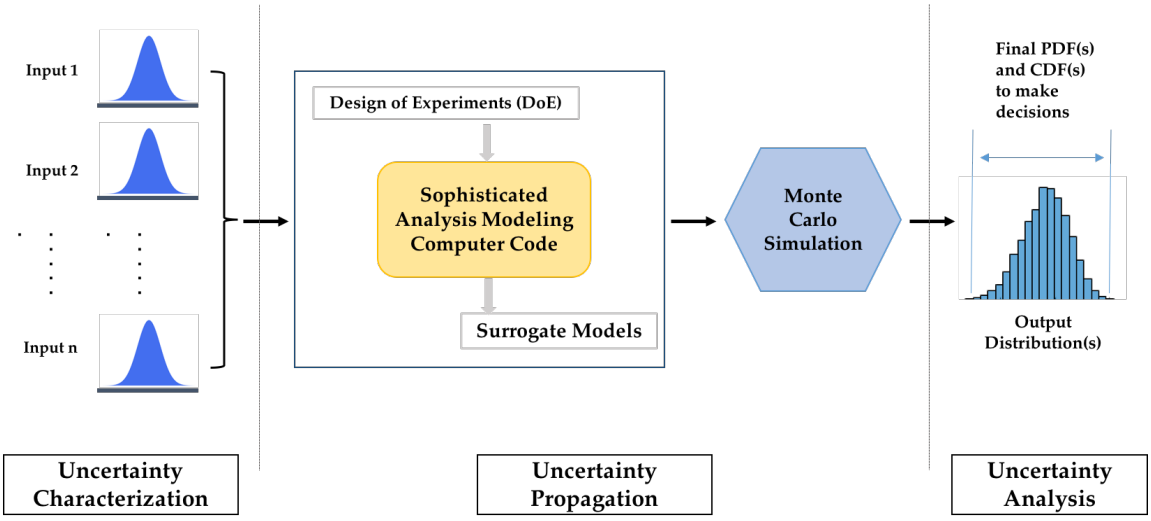


Figure 1.2: The MCS-based uncertainty propagation method

the dominant features of a complex model [5] (physics-based relationships between inputs and outputs can be retained). They have the potential to speed up the uncertainty propagation process without sacrificing accuracy or detail. The surrogate modeling techniques are mainly in three categories: data-driven, projection, and hierarchical-based. Methods in design of experiments (DoE) can provide data for data-driven surrogates, which approximate the original model through an empirical model that captures the input-output mapping. Two most commonly used surrogate models are response surface equations (RSE) and artificial neural network (ANN).

1.1.3 The Reality of Small Datasets

Uncertainty characterization is the first and an important step in the process of uncertainty quantification. Typically, to better characterize the distribution of an uncertain input, plenty of data is needed (say, with the dataset size $n > 100$). However in reality, we may not be able to collect *enough data*, or *complete information*, so that characterization of uncertain sources becomes difficult, for the following reasons:

1. **Cost of collecting data is high (money and/or time)**: for example, in industry, experiments for collecting one data point costs $5,000. With a limited budget ($100,000 for experiments), only 20 data points can be collected. Also, the time required for collecting the complete dataset may be too long, making it an impossible task within the timeframe of a project.

2. **Loss of information/unable to collect information**: for example, we want to collect aircraft weights for all the aircraft since 1950. Due to reasons such as loss of product manuals or lack of credibility for some data sources, we are only able to acquire half of the data needed.

3. **A complete dataset can still be small**: the whole population contains only 20-50 data, making it difficult to be accommodated by classical distributions

One commonly used way to propagate uncertainty from uncertainty inputs based on limited information and/or knowledge is the use of some "lack of knowledge distributions" [6], such as Triangular distribution and Student's t-distribution. Nevertheless, if they actually can not properly represent the uncertainty inputs, these distributions themselves may introduce additional uncertainty into the propagation process, making uncertainty in outputs even more difficult to be quantified. Some statistical methods, like the principle of maximum entropy, lead to the selection of a probability density function that is consistent with our knowledge and introduces

no unwarranted information [7]. Then is there a better way to better propagate uncertainty using the best probability distribution that reflects our current knowledge (staying loyal to the limited dataset)?

1.2 Background and Literature Review

1.2.1 Imprecise Probability

Imprecise Probability stands for uncertain probability distributions [8]. Quantification of uncertainty is mostly done by the use of precise probabilities: for each event A, a single (classical, precise) probability $P(A)$ is used, typically satisfying the Kolmogorovs axioms [9].

A limitation in this process is that classical probability requires a very high level of precision and consistency of information, such that sometimes the quality of underlying knowledge cannot be adequately represented using a single probability measure. One common flaw in the aerospace-related applications of probability is the arbitrary or unjustified use of distributions [10]. Uncertain factors can be modeled by a great variety of distributions (such as Triangular, Uniform, Beta, Weibull, and others), and encoded by experts through structured interviews. However, due to a lack of knowledge and/or resources, the normal distribution is often used as a default choice in many applications. In different cases, the incorrect assumptions of distribution types for random variables can lead to inaccuracies in uncertainty quantification process, and cause poor decisions [10].

This observation reveals the need for flexibility in accommodating distributions for a random variable. Imprecise probability provides important new methods that promise greater flexibility for uncertainty quantification as it: 1. Base inferences on weaker assumptions than needed for precise probabilistic methods; 2. Does not require experts to represent their judgments through a full probability distribution [9].

A series of influential developments in imprecise probability include Bayesian ap-

proaches, Interval methods, Dempster-Shafer evidence theory, and Fuzzy theory [8].

1.2.2 Literature Review: Latest Progress in Academia

In recent years, some researches have been conducted regarding the quantification and propagation of imprecise probabilities resulting from sparse and imprecise data, interval data, or small datasets [11] [12] [13] [14]. The researchers currently in this field are mainly from the community of mechanical engineering, civil engineering, and industrial engineering where reliability engineering matters. **The motivation for this research area is that both the frequentist and Bayesian statistical theories are well-suited for problems with large sample size, which is rarely available for many engineering applications** [11].

A parametric-based formulation has been developed and deemed relatively mature and rigorous, because this probabilistic approach divides total epistemic uncertainty created by small sample size into two parts: model-form uncertainty and parameter uncertainty. Details for the model-form uncertainty and parameter uncertainty are introduced in the next section. There are two dimensions of uncertainty [15]:

1. **Aleatory Uncertainty**: uncertainty involving unknown outcomes that can differ each time one runs an experiment under similar conditions

2. **Epistemic Uncertainty**: uncertainty involving missing knowledge concerning a fact that either is or is not true

Latest progress by Zhang and Shields in 2017 developed an efficient and parametric-based approach to propagate imprecise probabilities resulting from small datasets with the flow chart shown in Figure 1.3.

Main statistical elements in the parametric-based approach are introduced below:

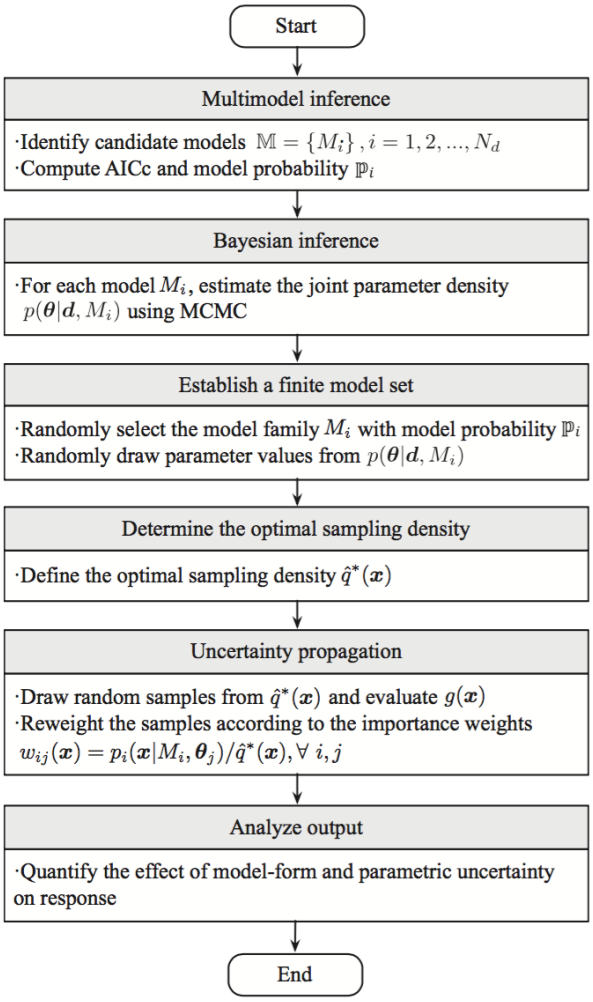


Figure 1.3: Flowchart of the proposed method for propagation of imprecise probabilities [11]

Candidate Models and Multimodal Inference

Starting with a small dataset, the first step for the researchers is to select a group os candidate distributions based on some knowledge on the presumed characteristics of the true distribution. For example, in Zhang and Shields' research [11], with 10 randomly generated data from a Lognormal distribution, they went on to select a group of 10 candidate parametric distributions: Inverse Gaussian Distribution, Lognormal Distribution, Gamma Distribution, Log-logistic Distribution, Rayleigh Distribution, Nakagami Distribution, Weibull Distribution, Levy Distribution, Exponential Distri-

bution, and F Distribution.

For each identified candidate distribution, a model selection method, corrected Akaike Information Criterion (AIC_c) is then used to determine the relative probability of each candidate distribution (probability for model p_i) [11]

$$AIC_c = -2\log(L(\hat{\theta}|\mathbf{d},\mathbf{M})) + 2K + \frac{2K(K+1)}{n-K-1}, \frac{n}{K} < 40 \tag{1.1}$$

$$\Delta_A^{(i)} = AIC_c^{(i)} - AIC_c^{min} \tag{1.2}$$

$$L(M_i|\mathbf{d}) = \exp(-\frac{\Delta_A^{(i)}}{2}) \tag{1.3}$$

$$p_i = p(M_i|\mathbf{d}) = \frac{\exp(-\frac{\Delta_A^{(i)}}{2})}{\sum_{i=1}^{N}\exp(-\frac{\Delta_A^{(i)}}{2})} \tag{1.4}$$

where K is the dimension of the parameter vector θ, n is the sample size of the dataset, and $L(\hat{\theta}|\mathbf{d},\mathbf{M})$ is the likelihood function given the maximum likelihood estimate of the parameters $\hat{\theta}$.

Equation 1.4 gives a model probability for each of the candidate distributions. As the end of the multi-model inference, a table was generated that ranks the relative probabilities of all the candidate models, as shown in Table 1.1.

Table 1.1: Table for the AIC_c model probabilities and the ranking [16]

Rank	**Distributions**	AIC_c	**Delta**	**Probability**
1	Inverse Gaussian	61.62	0.00	0.185
2	Lognormal	61.75	0.14	0.173
3	Gamma	61.94	0.34	0.156
4	Log-logistic	62.28	0.66	0.133
5	Rayleigh	62.35	0.74	0.128
6	Nakagami	62.38	0.77	0.126
7	Weibull	62.96	1.34	0.095
8	Levy	69.95	8.34	0.003
9	Exponential	72.75	11.13	0.001
10	F	98.39	36.77	0.000

In this example, 10 candidate distributions are ranked according to their model

probabilities (last column in Table 1.1). The inverse Gaussian distribution is the most likely model among all the candidate distributions with a probability of 18.5% (0.185), followed by Lognormal distribution and Gamma distribution. Before proceeding to the next step, some least likely candidate distributions whose probabilities are extremely low are eliminated. In this example, the three least likely candidates (Levy, Exponential, and F) are eliminated (colored in red in Table 1.1).

Bayesian Inference

For each candidate model survived from the multimodal inference, the Bayesian inference is performed to determine parameter uncertainty. For a specific model form M, model parameters become the next uncertain object to quantify. In this process, the PDF of the parameters are updated based on data collected, and a posterior distribution $p^*(\theta|\mathbf{d}, M)$ for the parameter is given by

$$p^*(\theta|\mathbf{d}, M) = \frac{p(\mathbf{d}|\theta, M)p(\theta; M)}{p(\mathbf{d}; M)} \propto L(\theta|\mathbf{d}, M)p(\theta; M) \tag{1.5}$$

$$p(\mathbf{d}; M) = \int L(\theta|\mathbf{d}, M)p(\theta; M)d\theta \tag{1.6}$$

where θ is the model parameter vector, $\mathbf{d}$ is the given data, M_i is the model form, and $p(\theta; M)$ is the prior distribution. Consequently, the Markov Chain Monte Carlo (MCMC) posterior joint parameter densities for all the candidate probability models are generated, as shown in Figure 1.4 for the same example.

Importance Sampling

Importance sampling is a variance reduction method in Monte Carlo integration, and the main idea is using samples generated from a different distribution rather than the distribution given [17]. Suppose $g(x)$ is another PDF, we have

$$\int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx \tag{1.7}$$

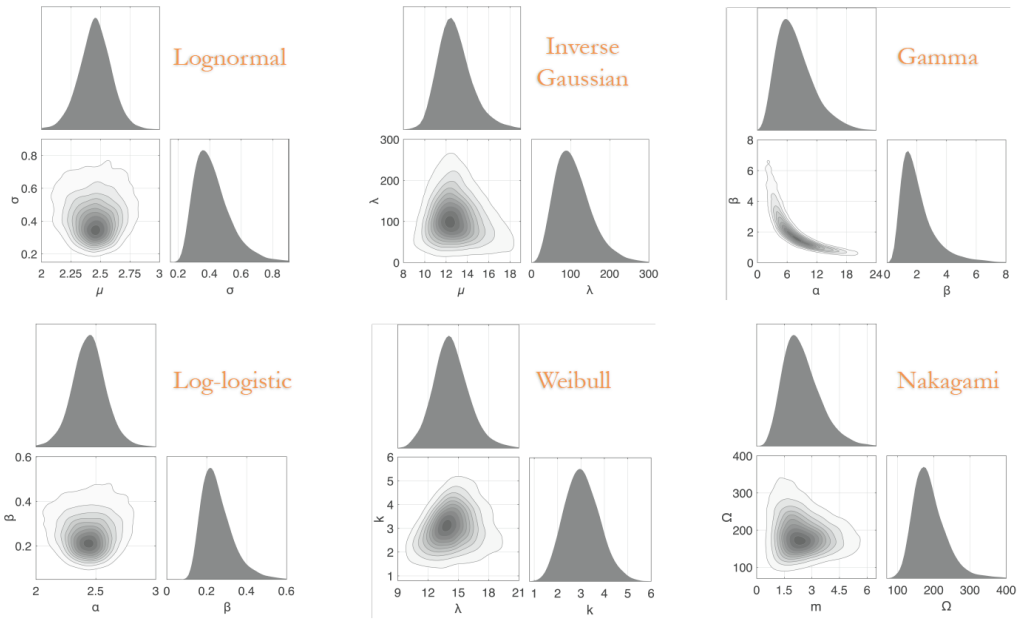


Figure 1.4: Example: MCMC posterior joint parameter densities for six of the candidate models [16]

where $\frac{f(x)}{g(x)}$ is the likelihood ratio, and we can change the function to be integrated as $h(x) \to h(x)\frac{f(x)}{g(x)}$, and change the PDF from $f(x)$ to $g(x)$. Based on this, now we can sample $X_1, X_2, \ldots, X_n$, i.i.d. from $g(x)$ and estimate the integration after.

Another important aspect is that we can choose $g(x)$ smartly so that the estimator has smaller variance than the one based on direct sampling. And if we want to show $\text{var}\{\hat{I}_2\} \leq \text{var}\{\hat{I}_1\}$, since $\text{E}\{\hat{I}_2\} = \text{E}\{\hat{I}_1\}$ (unbiased estimator), that is equivalent to $\text{E}\{\hat{I}_2^2\} \leq \text{E}\{\hat{I}_1^2\}$. From the relationship

$$\int [h(x)\frac{f(x)}{g(x)}]^2 g(x)dx \leq \int h(x)^2 f(x)dx \tag{1.8}$$

using the Cauchy Schwartz Theorem, we finally arrive at $g(x) = \frac{|h(x)|f(x)}{\int |h(v)|f(v)dv}$.

Optimal Sampling Density for Multiple Distributions

Since the target density is unknown, and the parametric-based formulation involve multiple plausible probability models M_i, each with uncertain parameters θ quantified

through Bayesian inference, we aim to identify a single sampling density $q^*(x)$ that is representative of all $p_i(\mathbf{x}|\theta)$ and can be used in importance sampling to propagate the full set of probability models.

$$\begin{aligned} \mathrm{E}[\mathcal{M}(\mathbb{M}||Q)] = \sum_{i=1}^{N_d} E_\theta[\mathcal{M}(M_i||Q)] &= \sum_{i=1}^{N_d} \frac{1}{2} E_\theta\big[\int_\Omega (p_i(\mathbf{x}|\theta) - q(\mathbf{x}))^2 d\mathbf{x}\big] \\ &= E_\theta\big[\int_\Omega \sum_{i=1}^{N_d} \frac{1}{2} (p_i(\mathbf{x}|\theta) - q(\mathbf{x}))^2 d\mathbf{x}\big] \end{aligned} \tag{1.9}$$

And an optimization problem is defined as

$$\begin{aligned} &\text{Minimize} \quad E_\theta\big[\int_\Omega \sum_{i=1}^{N_d} \frac{1}{2} (p_i(\mathbf{x}|\theta) - q(\mathbf{x}))^2 d\mathbf{x}\big] \\ &\text{Subject to} \quad \int_\Omega q(\mathbf{x}) d\mathbf{x} - 1 = 0 \end{aligned} \tag{1.10}$$

Solving for $q(\mathbf{x})$ gives the analytical solution of the minimizer

$$\hat{q}^*(\mathbf{x}) = \frac{1}{N_d} \sum_{i=1}^{N_d} E_\theta[p_i(\mathbf{x}|\theta)] = \sum_{i=1}^{N_d} \mathbb{P}_i \mathbb{E}_\theta[p_i(\mathbf{x}|\theta)] \tag{1.11}$$

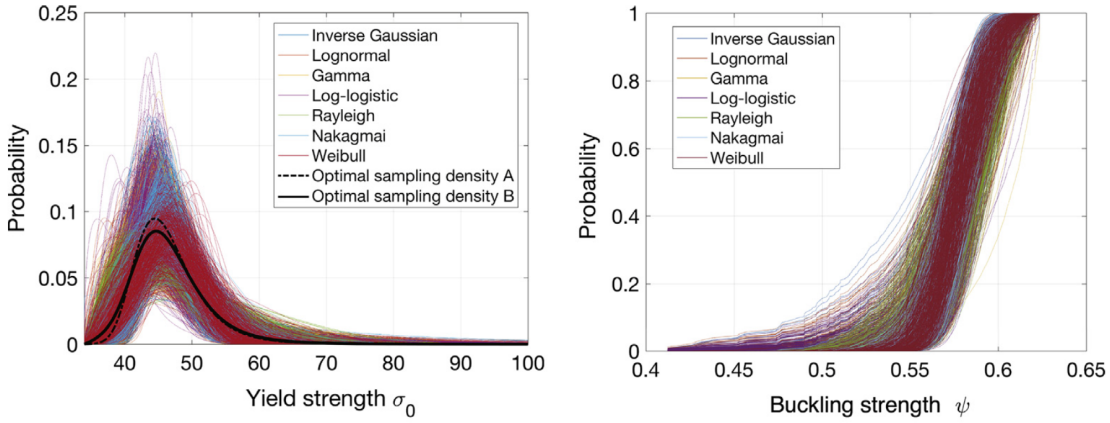
where $\mathbb{P}_i$ is the AIC model probability that satisfies $\sum_{i=1}^{N_d} \mathbb{P}_i = 1$.

Uncertainty Propagation

The uncertainty propagation under this parametric-based formulation is conducted through importance sampling with the sampling density $\hat{q}^*(\mathbf{x})$. Samples are drawn from $\hat{q}^*(\mathbf{x})$ and re-weighted according to importance weights.

For each sample, the parametric distribution is randomly drawn according to AIC probabilities $\mathbb{P}_i$ (Table 1.1), and the parameters of that distribution are then randomly drawn from their joint parameter densities (Figure 1.4). When the process is repeated for 5,000 times, 5,000 candidate densities are generated, with two optimal densities shown in Figure 1.5.

Figure 1.6 and 1.7 show the evolution of candidate target densities and CDFs for the output for increasing data sizes. Figure 1.6 shows that the cloud of candidate densities narrows when data are gradually added, but loss of optimality for the orig-



(a) (b)

Figure 1.5: (a) Candidate PDFs and the optimal sampling densities from ten yield stress values, and (b) collection of candidate empirical CDFs for the output [11]

inal optimal sampling density may happen, even if the density can still be used for propagation. Figure 1.7 shows the corresponding set of CDFs for the output (the system model in this example only has one input and one output). As we gradually add more data in, the cloud of output CDFs also narrows towards the true CDF as expected. The author of this thesis thinks that the could of output CDFs shown in Figure 1.7 can also be used to quantify the uncertainty of probability at each x under this formulation in the form of point-wise confidence interval.

1.2.3 Model-Form Uncertainty and Parameter Uncertainty

Most of the research efforts in this field separate model-form and parameter uncertainties [11] [12] [13] [14]. Researches that delineated the effects of model-form and parameter uncertainties provide more insights in these two types of uncertainties for parametric distributions. Definitions for model-form and parameter uncertainties are:

- **Parameter Uncertainty**: when only finite data is available, and a distribution type is assured for the dataset, there is uncertainty associated with the parameter estimates. The importance of parameter uncertainty increases when data is sparse or imprecise. It is one example of epistemic uncertainty.

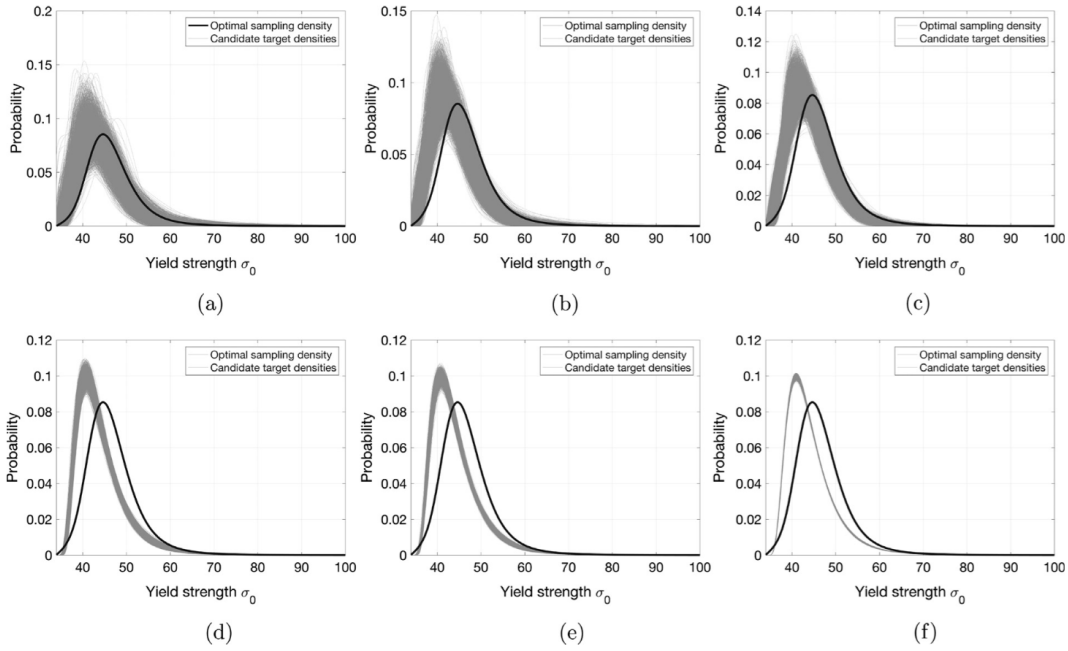


Figure 1.6: Optimal sampling density with candidate target densities based on: (a) 25 data, (b) 50 data, (c) 100 data, (d) 500 data, (e) 1000 data, and (f) 10,000 data [11]

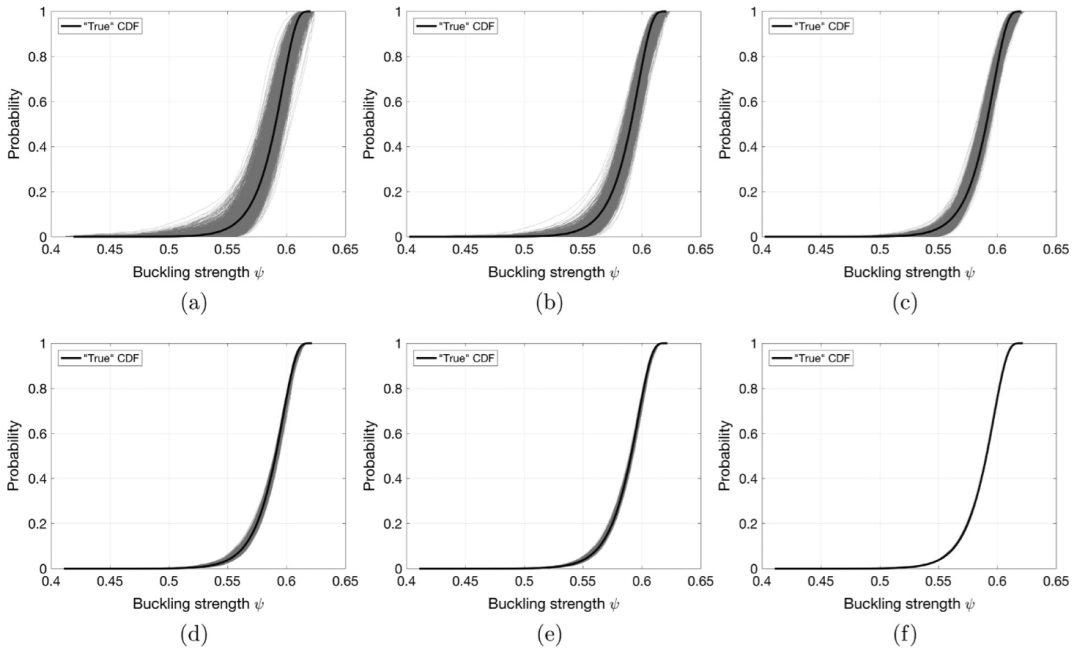


Figure 1.7: CDFs for the output based on: (a) 25 data, (b) 50 data, (c) 100 data, (d) 500 data, (e) 1000 data, and (f) 10,000 data [11]

- **Model-form Uncertainty**: prior to the estimation of distribution parameters, the choice of distribution type is another example of epistemic uncertainty. This is because the true underlying distribution type may not be possible to identify due to the available sparse and imprecise data [14].

Research by Zhang and Shields in 2017 shows that in their study case, the parameter uncertainties are dominant because they accounts for the greater proportion of the total uncertainty. However in the meantime, the model-form uncertainties are important as well [11]. In the same study case, uncertainty propagation results show that the parameter uncertainty diminishes more slowly than the model-form uncertainty and can remain noticeable even for very large datasets.

1.2.4 Notable Limitations for Parametric-based Methods

A combination of several classical statistical methods, the parametric approach on the propagation of imprecise probabilities due to small datasets is rigorous and valuable. Nevertheless, such a formulation still has some notable limitations, which make it impossible to be applied to a broader range of problems:

1. **Selection of Candidate Probability Models**: as the first step of the formulation, it requires people (researchers, experts, etc.) to manually supply a group of candidate probability models based on limited knowledge and judgments. Ideally, given a sufficiently diverse set of candidate models, the set of models may be sufficient to span the epistemic uncertainty, but this can not be proved. Even if the set of candidate models actually include all the parametric models within certain range (such as $[0, +\infty)$), the whole process would be computationally too expensive.

2. **Constraint of Parametric Models**: another limitation is that the formulation, and the set of candidate models is constrained to well-established para-

metric models (such as Normal, Lognormal, Weibull, etc.). Although it is true that those well-known parametric models are very useful in modeling numerous events that happen in the nature, they can only be used to model a limited range of things. Another thing related to this is that most of the parametric models are unimodal, which limits the parametric-based approach from estimating multimodal distributions (distributions with more than one peak) without using mixture models.

3. **Loss of Optimality**: loss of optimality for the optimal parametric distribution may happen as more data are added in. The parametric-based formulation uses an optimization set-up, combined with multi-modal inferences to obtain a single optimal sampling density. Yet later when the Monte Carlo simulation is used and a family of multiple candidate PDFs is obtained, loss of optimality in the importance sampling density may happen.

1.3 Problems Related to the Aviation Environmental Design Tool (AEDT)

1.3.1 AEDT Introduction

Aviation Environmental Design Tool (AEDT) is a software system that models aircraft performance in space and time to estimate fuel consumption, emissions, noise, and air quality consequences [18]. A primary objective of AEDT is to help the analyst efficiently answer questions of interest about the environmental consequences of aviation activities. It is a comprehensive tool that provides information to FAA stakeholders on each of these specific environmental impacts. Main components of the AEDT software system are shown in Figure 1.8.

The AEDT has been used in a number of global and US policy making processes, including International Civil Aviation Organization (ICAO) standards for Noise, and emissions (CO_2, NO_x).

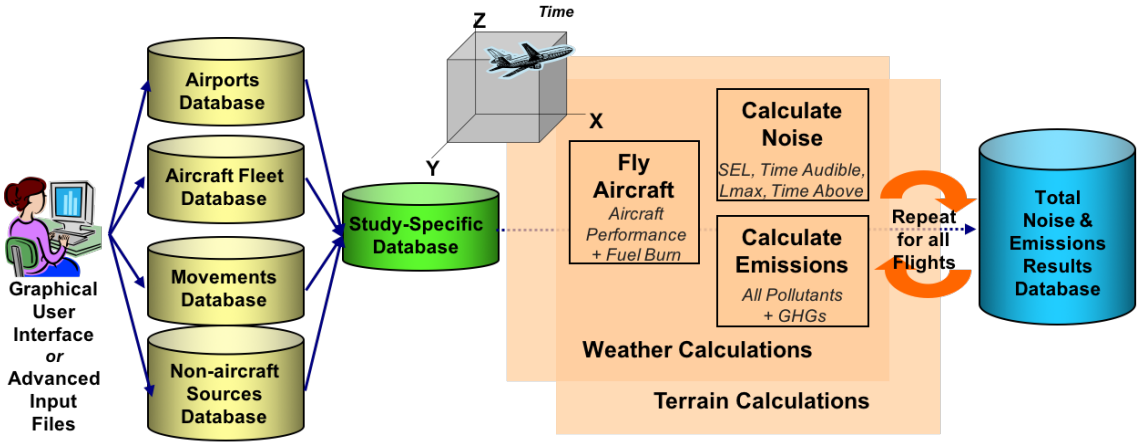


Figure 1.8: Main components of the AEDT software system

1.3.2 Difficulty to Acquire Complete Data

Another challenge for the researchers in AEDT is the difficulty to acquire complete datasets for all the input parameters, and the difficulty comes from two aspects:

1. **The number of input parameters**: input parameters of the AEDT are mainly under two categories: airport atmosphere (such as temperature, pressure, etc) and aircraft performance (such as thrust, weight, etc). Overall, there are more than 60 input parameters under airport atmosphere (5 inputs) and aircraft performance (55+ inputs).

2. **Incomplete data for some parameters**: the researchers have encountered situations that the aircraft performance input data do not cover all the unique aircraft models, which is another potential source of uncertainty. Other than that, the incomplete data may come from other dimensions, such as time, airline, area, airport, and so on.

Therefore, for a complex model like the AEDT which has a great number of input parameters, it is always not easy to collect complete datasets as needed. As a result, incomplete data/small datasets are very likely to happen, which makes the studies more challenging.

1.4 Research Questions and Hypotheses

Motivated by the reality of small datasets in research projects in both academia and industry, and the limitations of parametric-based approach to propagate imprecise probabilities stated above, the thesis work tries to develop a nonparametric-based approach to propagate imprecise probability due to small datasets. This thesis work consists of two main parts: Part I provides nonparametric density estimations on the one-dimensional data; Part II generates two-dimensional density estimations, which utilizes copulas to capture the correlation between variables by linking the marginal densities.

The research questions outlined here are established as means to guide the overall research plan, and the corresponding hypotheses will then be tested through a series of experimentations. The first research question is an overarching research question of this thesis which addresses the overall objective.

> **Overarching Research Question**: *What is an appropriate method to propagate uncertainty of imprecise probabilities due to small dataset while solving the constraints and limitations of the current parametric-based method and staying consistent with our knowledge without introducing unwarranted information?*
>
> **Overarching Hypothesis**: *If the nonparametric statistical elements are properly used, a nonparametric-based method facilitates the quantification and propagation of uncertainty created by lack of sufficient statistical data, and can be used in a broader range than the parametric-based method.*

The research questions 1 to 4 break down the overall research question into more detailed parts, such that the overall research objective can be obtained through research from different steps. The first research question seek to identify an optimal density estimation based on small dataset.

Research Question 1: *How can the nonparametric density estimation methods be used to obtain an optimal sampling density to estimate the true density using small dataset and to propagate uncertainty? The key part in the estimation is to identify a nonparametric model without overfitting or underfitting.*

Hypothesis 1: *Such an optimal density estimation can be done through the nonparametric Kernel Density Estimation (KDE) with an optimal bandwidth selection method.*

Since the most important task for propagating the uncertainty of imprecise probabilities is to obtain an optimal sampling density based on small dataset, the first and foremost action item is to obtain such a density completely in a nonparametric way. The Kernel Density Estimation (KDE) is the most appropriate nonparametric way to estimate the probability density function of a random variable. Since the most challenging part in the KDE is the identification of an appropriate bandwidth such that the density estimated is neither overfitted nor underfitted. Although this is a challenging area in the theory of statistics, a persuasive bandwidth solution must be constructed for the optimal sampling density.

Research Question 2: *Apart from the optimal sampling density, can we provide another more conservative sampling density such that more potential extreme values in the output distributions can be captured through uncertainty propagation process? The necessity for this option exists because uncertainty is an inherent nature of small datasets.*

Hypothesis 2: *Such a density estimation with larger variance than the optimal sampling density can be obtained through the appropriate use of nonparametric statistical elements such as KDE and bootstrap.*

The research question 2 is a crucial study of this thesis research, because for the case of lacking sufficient statistical data, uncertainty will exist anyway. Although the

optimal sampling density provides a dimension to tackle this challenge, it does not reflect risk. What is needed here is another sampling density estimation that contains more uncertainty than the optimal sampling density. By propagating this dimension of sampling density, the target is to simulate more potential extreme values for the outputs after uncertainty propagation.

> **Research Question 3**: *When dependencies among the variables exist, how can the nonparametric-based method be adapted to capture the dependencies and generate multi-dimensional density estimations?*
>
> **Hypothesis 3**: *The use of Copulas and the Sklar's Theorem can capture the dependencies between the variables and link the marginal nonparametric densities to create a joint density.*

The third research question comes from a real need in complex systems model. In such complex models with a large number of input variables, dependencies may exist among some of the input variables. If the dependencies are ignored and all the input variables are treated as independent, uncertainty in the outputs can be either overestimated or underestimated. Therefore, the multi-dimensional solution (two-dimensional is the focus of this thesis) is highly necessary to be created to capture those dependencies such that in uncertainty propagation, random samples can be generated from the joint sampling densities for dependent variables.

> **Research Question 4**: *How can the linear correlation between two variables be identified and confirmed based on small dataset?*
>
> **Hypothesis 4**: *The linear correlation between two variables can be identified through a combination of curvilinear regression, correlation coefficient, and different types of confidence intervals.*

The last research question is brought out by the process of identifying and confirming linear correlation between two variables. This step is important because Copulas and Sklar's Theorem will be used only when the linear relationship between the two variables (prerequisites for Copulas) is confirmed. And in the small dataset case, such a judgment is more challenging because of the uncertainty inherent with incomplete information. Before applying Copulas and Sklar's Theorem to model the dependencies, we must first be confident that the linear correlation do exist, based only on the small dataset that we have. A decision tree that involves different judgment methods, such as multiple curvilinear regressions, the correlation coefficient, and different confidence intervals for the correlation coefficient is needed.

This thesis research is conducted following the same sequence of the 4 research questions above. In the end, all the research questions will be answered, based on the effectiveness of different models built in the process.

CHAPTER 2
1-D NONPARAMETRIC DENSITY ESTIMATION

2.1 Parametric vs. Nonparametric Statistics

Since the term "Nonparametric" is the main feature that constructs this thesis work and distinguishes this work from the literatures, it is necessary to briefly introduce nonparametric statistics and the differences between parametric and nonparametric statistics.

- **Parametric Statistics**: a branch of statistics with the assumption that the sample data comes from a population that follows one of the *parametric distributions* - probability distributions that are based on a fixed set of parameters [19]. In fact, most of the well-known statistical methods and distributions are parametric [20]. For example, the normal (Gaussian) distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ is defined by two parameters: the mean (μ), and the variance (σ^2). Other common parametric distributions include Lognormal Distribution $X \sim Lognormal(\mu, \sigma^2)$, Inverse Gaussian Distribution $X \sim IG(\mu, \lambda)$, Gamma Distribution $X \sim \Gamma(\alpha, \beta)$, Exponential Distribution $X \sim Exp(\lambda)$, etc.

- **Nonparametric Statistics**: the term "Nonparametric" was first mentioned by Jacob Wolfowitz [21]. And nonparametric statistics is broadly defined to include all methodology that does not use a model based on a single parametric family. For nonparametric distributions, some representative ones include Histogram, Empirical Distribution and Kernel Distribution. The Kernel Density Estimation (KDE) is a crucial element in this thesis work.

When comparing parametric and nonparametric statistics, a key difference is that

nonparametric methods are not based on parametric assumptions - in the nonparametric case, functional forms of the distributions are unknown. Apparently, nonparametric statistics has its own advantages, which make it a better choice under some circumstances. Sometimes, data generated from complex experiments and messy sampling plans may not be attributed to any well-known parametric distributions. In those cases, making parametric assumptions when people are not sure about the underlying distribution of the data is not reasonable. Nonparametric methods can be applied regardless of the true distribution of the data, and are also called "distribution-free methods" [21].

2.2 Histogram

Although histogram is not included as a part in estimating the densities in this work, it is the simplest nonparametric density estimator [22]. Although it is very old-fashioned, histograms are easy to draw and are very widely used in applied work.

Let X_1, X_2, ..., X_n be independent random variables with common density f, the empirical histogram for the X's is often used to estimate f. Let h be the cell width on the interval where data fall, and let N_j be the number of data falling in the jth class interval $[x_0 + jh, x_0 + (j+1)h]$, then the height of the histogram $H(x)$ on this interval is defined as N_j/kh. This definition forces the area under H to be 1 [23].

An standard measure of discrepancy is the mean square difference

$$\delta^2 = \mathbb{E}\{\int_I [H(x) - f(x)]^2 dx\} \tag{2.1}$$

The histogram can become a valuable estimation when the discrepancy δ^2 is made small, and that can be obtained by choosing a cell width h which minimizes δ^2. The selection of h is the most crucial part in histogram; relevant methods and discussions can be found in literatures [23].

2.3 Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a nonparametric representation of the probability density function (PDF) of a random variable [24]. It can be used when a parametric distribution cannot properly describe the data, or when people want to avoid making assumptions about the distribution of the data.

The KDE is defined by a smoothing function and a bandwidth value. It sums the component smoothing functions for each data value to produce a smooth, continuous probability curve, as shown in Figure 2.1.

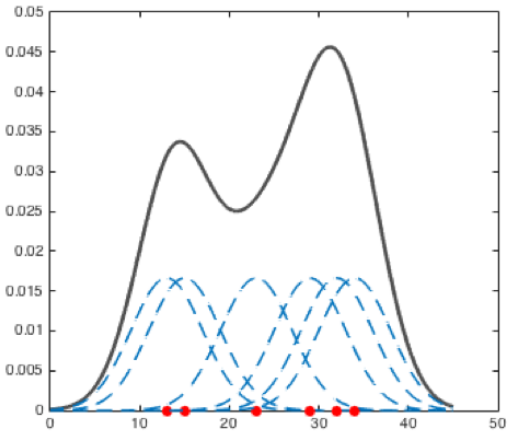


Figure 2.1: Schematic of the KDE: data (red dot), kernel base function (blue), and kernel density estimation(gray)

With a sample X_1, X_2, ..., X_n, we form the density estimator as [21]

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{2.2}$$

For $X_i = x_i$, $i = 1, 2, ..., n$, the kernel function K is generally chosen to be a unimodal probability density symmetric about zero, and satisfies $\int K(x)dx = 1$. The smoothing parameter h is known as the bandwidth [25].

The kernel function K has four main basic kernel functions: normal, epanechnikov, box, and triangle, as shown in Figure 2.2. The normal kernel function is used throughout the one-dimensional part of this thesis work.

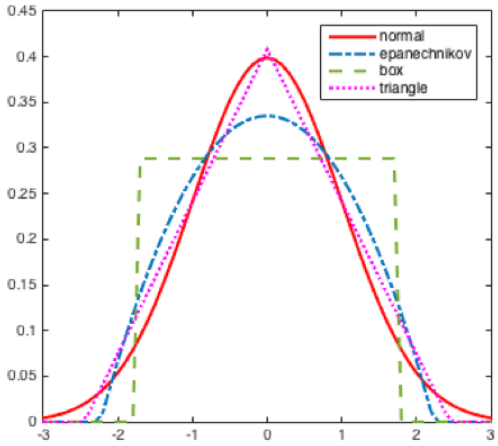


Figure 2.2: The four basic kernel functions: normal, epanechnikov, box, triangle

2.4 Bandwidth Selection

Bandwidth selection is the most difficult and significant part for KDE, and have been widely studied over the past decades. A key question here is: how to select an appropriate bandwidth value such that the kernel density estimated through small dataset is neither overfitting nor underfitting? Normally, a kernel density with different bandwidth values (h) can be [26]:

- **Under smoothed**: an estimator with high variability *[h is too small]*
- **Over smoothed**: too smooth to obscure the underlying structure *[h is to large]*
- **Optimally smoothed**: the estimation is close to the true density

Figure 2.3 demonstrates the effect of different bandwidths used in KDE on the same sample. It can be observed that when the kernel density is optimally smoothed, the shape of the density is the closest to the true density, and can therefore provide a good reference in the uncertainty propagation process.

To find an optimal bandwidth for the density estimator, the starting point is a concept called the Mean Integrated Squared Error (MISE). It is the discrepancy

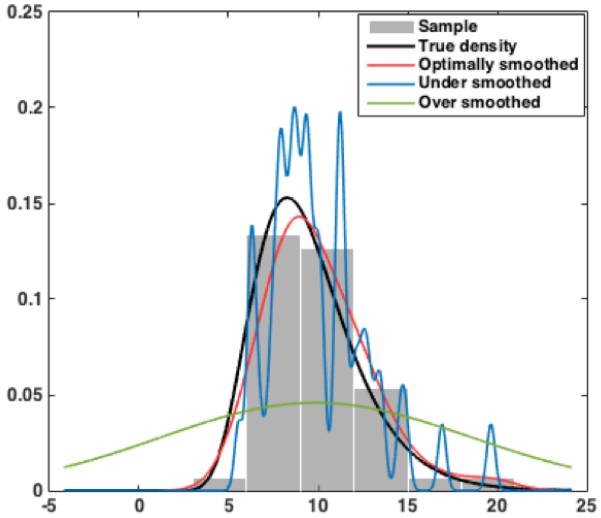


Figure 2.3: Effect of different bandwidths on the shape of kernel density (compared to true density)

between $\hat{f}$ and f, and is given by [21]

$$\begin{aligned} \text{MISE}(\hat{f}_h) &= \mathbb{E}\Big\{ \int (f(x) - \hat{f}_h(x))^2 dx \Big\} \\ &= \int \text{Bias}^2(\hat{f}_h(x)) dx + \int \text{Var}(\hat{f}_h(x)) dx \end{aligned} \tag{2.3}$$

Assuming that the true density $f(x)$ is continuous and smooth, and that the kernel has finite fourth moment, it can be shown that [25]

$$\text{Bias}^2\{\hat{f}(x)\} = \frac{h^2}{2} u_2(K) f''(x) + O(h^2) \tag{2.4}$$

$$\text{Var}\{\hat{f}(x)\} = \frac{1}{nh} R(K) f(x) + O\big(\frac{1}{nh}\big) \tag{2.5}$$

where

$$R(K) = \int K^2(x) dx \tag{2.6}$$

$$u_2(K) = \int x^2 K^2(x) dx > 0 \tag{2.7}$$

when choosing the normal (Gaussian) kernel for K,

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) \tag{2.8}$$

By adding the leading variance and squared bias terms appeared in Equations 2.4 and 2.5, we get the Asymptotic Mean Squared Error (AMSE) [25]

$$\text{AMSE}\{\hat{f}_h(x)\} = \frac{1}{nh} R(K) f(x) + \frac{h^4}{4} u_2(K)^2 [f''(x)]^2 \tag{2.9}$$

Assume the integrability on f, it gives us the Asymptotic Mean Integrated Squared Error (AMISE) by integrating AMSE

$$\text{AMISE}\{\hat{f}_h(x)\} = \frac{1}{nh} R(K) + \frac{h^4}{4} u_2(K)^2 R(f'') \tag{2.10}$$

where

$$R(f'') = \int [f''(x)]^2 dx \tag{2.11}$$

Finally, the bandwidth value that minimizes the AMISE is our optimal choice for the density estimator, given by

$$h_{\text{AMISE}} = \Big[\frac{R(K)}{u_2(K)^2 R(f'')}\Big]^{1/5} n^{-1/5} \tag{2.12}$$

2.4.1 Rules of Thumb

The rules of thumb methods for choosing a global bandwidth h are computationally the simplest [25]. It is based on replacing $R(f'')$, the only unknown part for h_{AMISE}, by its value for a parametric family. In general, the rules of thumb are trying to build the relationship between the optimal h and two important characteristics of a dataset: sample size (n), and sample standard deviation (S).

Assuming that the underlying distribution is normal, Silverman (1986) showed that the resulting bandwidth in Equation 2.12 becomes

$$h_{\text{AMISE}_{\text{NORMAL}}} = 1.06 \sigma n^{-1/5} \tag{2.13}$$

Later Jones et al. (1996) studied the Monte Carlo performance of the $h_{\text{AMISE}_{\text{NORMAL}}}$ based on standard deviation, and considered

$$h_{\text{SNR}} = 1.06Sn^{-1/5} \tag{2.14}$$

where S is the sample standard deviation.

In order to not miss the bi-modality, Silverman (1986) recommended reducing the factor 1.06 in h_{SNR} to 0.9, and using the smaller of two scale estimates to replace the standard deviation [27]. This is the Silvermans reference bandwidth

$$h_{\text{SROT}} = 0.9An^{-1/5} \tag{2.15}$$

where $A = min\{S, IQR/1.34\}$

The last bandwidth selection method considered in this work was developed by Terrell and Scott (1985) [28] and Terrell (1990) [29]. They developed a bandwidth selection method based on the maximal smoothing principle to produce oversmoothed density estimates [25]. When using the standard normal kernel, the oversmoothed bandwidth is

$$h_{\text{OS}} = 1.144Sn^{-1/5} \tag{2.16}$$

All the three rules of thumb methods are used in the 1-D density estimations in this work. See section 2.4.4 for the detailed allocations of h_{SNR}, h_{SROT}, h_{OS}.

2.4.2 Cross-Validation Methods

Cross-validation is one of the classical methods for kernel density estimation bandwidth selection [30]. The idea of cross-validation is to use part of data for training, and the remaining data for estimating testing error [31]. When partitioning the original dataset, there are two types of common techniques:

1. **K-fold (K-fold CV)**: For a number K, split the training samples into K batches, train on all but the kth part, and then validate on the kth part. Iter-

ating over $k = 1, \ldots, K$

2. **Leave-one-out (LOOCV)**: When $K = n$, leave out one data point at a time

The cross-validation methods start with the Integrated Squared Error (ISE), which measures the closeness of $\hat{f}$ and f

$$\begin{aligned} \mathrm{ISE}(\hat{f}_h) &= \int (f(x) - \hat{f}_h(x))^2 dx \\ &= \int (\hat{f}_h(x))^2 dx - 2\int \hat{f}_h(x) f(x) dx + \int f^2(x) dx \end{aligned} \tag{2.17}$$

in which, the term $\int (\hat{f}_h(x))^2 dx$ depends solely on the density estimate and can be evaluated numerically; the third term $\int f^2(x)dx$ does not involve h and can therefore be ignored. Bowman (1984) estimated the central term $\int \hat{f}_h(x) f(x) dx$ as

$$\frac{1}{n}\sum_{n=1}^{n} \int (\hat{f}_{-i}(y))^2 dy - \frac{2}{n}\sum_{n=1}^{n} \int \hat{f}_{-i}(X_i) \tag{2.18}$$

where $\hat{f}_{-i}(y)$ denotes the kernel estimator constructed with X_i deleted.

With Hall (1983) [32] showed that

$$\sum_{n=1}^{n} \int (\hat{f}_{-i}(y))^2 dy = \int (\hat{f}_h(y))^2 dy + O_p\Big(\frac{1}{n^2 h}\Big) \tag{2.19}$$

the final Least Squares Cross-Validation (LSCV) based criterion is [33]

$$\mathrm{LSCV}(h) = \int (\hat{f}_h(y))^2 dy + \frac{2}{n}\sum_{n=1}^{n} \int \hat{f}_{-i}(X_i) \tag{2.20}$$

This version is used by most of the researchers, and the resulting value of h that minimizes LSCV(h) is demoted by h_{LSCV}.

Apart from LSCV(h), some other cross-validation criteria include [30]:

$$\mathrm{LCV}(h) = -\sum_{n=1}^{n} \log \hat{f}_h(X_i) - \sum_{n=1}^{n} \log(1 - K(0)/(nh\hat{f}_h(X_i))) \tag{2.21}$$

$$\mathrm{AIC}(h) = -\sum_{n=1}^{n} \log \hat{f}_h(X_i) - \sum_{n=1}^{n} K(0)/(nh\hat{f}_h(X_i)) \tag{2.22}$$

With the LSCV criterion, the next step in cross-validation is to plot LSCV(h)

against h, and identify the h_{LSCV} that minimizes LSCV(h). An example of such plot is shown in Figure 2.4. For the test case in this Figure, $h_{\text{LSCV}} = 0.68$ is the minimizer of LSCV(h).

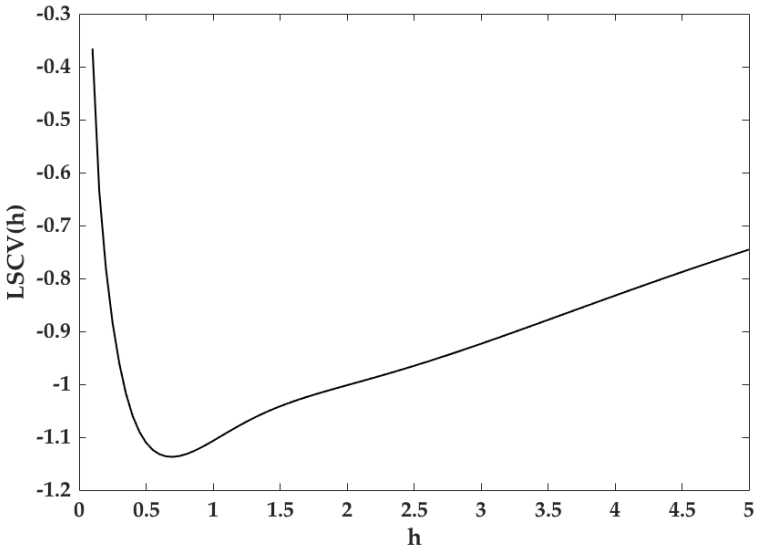


Figure 2.4: Plot of the least squares cross-validation function LSCV(h) against h

2.4.3 Plug-in Methods

Since LSCV has a slow rate of convergence, a category of faster converging methods called plug-in methods was invented [25]. It replaces the unknown quantity $R(f'')$ in h_{AMISE} with an estimate $R(\hat{f}_g'')$, and the bandwidth g chosen by the user is called "pilot estimate". The plug-in methods is recommended by some authors in literature, for its overall good performance [25]. Nevertheless, other authors (such as Loader (1999)) did their research and challenged the claimed superiority of plug-in methods on several fronts [30]. From where they stand, the plug-in methods could fail when the arbitrary specification of pilot bandwidth is wrong, and may result in inefficient estimates. The plug-in methods is not used in this thesis work, but may be considered as an option if supported by more evidence in the small dataset case and issues stated above can be controlled in an acceptable range.

2.4.4 Final Formulation: A combination of Rules of Thumb and Cross Validation

Taking into account all the three major categories of bandwidth selection methods for kernel density estimates (rules of thumb, cross-validation, and plug-in), the author of this thesis uses a combination of rules of thumb and cross-validation methods for choosing a bandwidth for the optimal one-dimensional nonparametric density estimate. Detailed assignment of the methods in the one-dimensional formulation can be found in Table 2.1. The meaning of all the usages can be found later in this chapter.

Table 2.1: Assignment of bandwidth selection methods through the one-dimensional formulation

Methods	**Details**	**Usage**
Rules of Thumb	h_{SNR}	Used when generating B nonparametric bootstrap family, for its simplicity in computation
Rules of Thumb	h_{SROT}	Used as the lower bound for cross-validation, for its property to include bi-modality
Rules of Thumb	h_{OS}	Used as the upper bound for cross-validation, for its over-smoothed estimates
Cross-Validation	COOCV	Used when the number of data is less than or equal to 100
Cross-Validation	10-fold CV	Used when the number of data is greater than 100

To further elucidate this assignment, there is one major distinction between rules of thumb and cross-validation methods - cross validation often brings variability and undersmoothing [30]. While this reflects the uncertainty of bandwidth selection, a small bandwidth value given by cross-validation is likely to cause overfitting in the small dataset case. Therefore, we can make use of the rules of thumb bandwidths to control the range of h in cross-validation, and thus limit overfitting or underfitting to a reasonable range.

Among the three rules of thumb bandwidths (h_{SNR}, h_{SROT}, h_{OS}), h_{SNR} is one that is closest to what we called "optimal estimate", while the other two have different targets. h_{SROT} is an attempt to not miss bimodality, and can therefore be used as the lower bound in bandwidth selection. Likewise, since h_{OS} is based on the maximal smoothing principle and produces oversmoothed density estimates, this oversmoothed bandwidth can be used as the upper bound in bandwidth selection. So before the cross-validation process get started, the setting first limits h in the range of $[h_{\text{SROT}},\ h_{\text{OS}}]$.

Next, we add the element of cross-validation, and plot LSCV(h) against h within $[h_{\text{SROT}},\ h_{\text{OS}}]$. The final optimal bandwidth selected is the h_{LSCV} that minimizes LSCV(h) within $[h_{\text{SROT}},\ h_{\text{OS}}]$. In this way, we can use the classical cross-validation method within a small range to prevent overfitting or underfitting while saving some computational costs.

In data partition, the Leave-one-out Cross-validation (LOOCV) method is used when the dataset size is less than or equal to 100. The reason is that the LOOCV does the best job in maximizing the usage of every data in small dataset, and can prevent the randomness of data partitioning from happening. Due to its computational costs with larger amount of data, the 10-fold Cross-validation (10-fold CV) is a better choice when n is greater than 100.

2.5 Bootstrap

Bootstrap is a statistical technique which means *random sampling with replacement* [21]. It learns about the sample characteristics by taking resamples and use the information to infer to the population characteristics. In this process, some of the n items from the original samples X_1, X_2, . . . , X_n can appear more than once. Bootstrap was first mentioned by Bradley Efron in 1979, and the theory behind it is sophisticated, being based on Edge-worth Expansions [34]. It is one of several controversial

techniques in statistics [21], but has been used a lot in quantifying uncertainty since it's introduction.

To further explain the process of random sampling with replacement, for example, if the original data sample is [1, 3, 5], then possible bootstrap samples with the same sample size can be [1, 1, 1], [3, 3, 5], [5, 3, 1], etc,.

A basic process of bootstrap can be found in Figure 2.5. It provides a powerful tool to calculate the standard error of an estimator, construct confidence intervals, and many other uses [34]. Because of its capability of calculating such intervals, when we only have small datasets for further use due to limitations in reality, bootstrap is a powerful tool to extract more information from the original datasets without generating additional data.

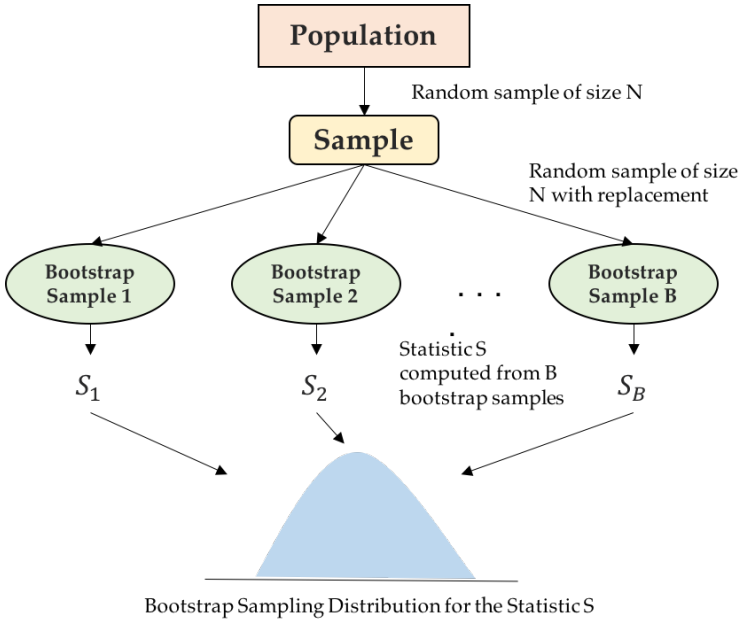


Figure 2.5: A basic bootstrap process for estimating the distribution of some statistic for population [34]

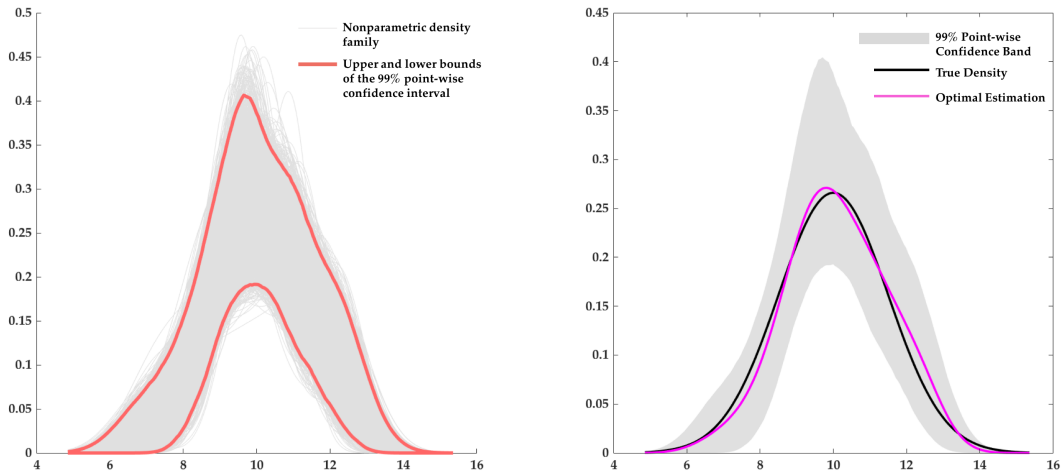
2.6 The 99% Point-wise Confidence Band

In this part, two key nonparametric statistical methods: kernel density estimation and bootstrap, are used to further infer the characteristics of the true density based

on small datasets. The objective of the one-dimensional steps going forward is to better tackle the challenge of uncertainty which is inherent in small datasets.

To begin with, we first generate a family of 10,000 nonparametric kernel densities. This can be done by the following steps:

1. Use bootstrap to obtain new sample x_b with same size
2. Estimate kernel density on x_b with normal base and bandwidth h_{SNR}
3. Repeat steps 1 and 2 for B (B = 10,000) times to generate a family of B nonparametric densities $\{\hat{f}_b : b = 1, ..., B\}$
4. Find the 99% point-wise confidence band among the nonparametric densities



(a) The NP density family and the 99% point-wise confidence band boundaries

(b) The 99% point-wise confidence band for the inference of the shape of the true density

Figure 2.6: Schematic of the nonparametric density family, 99% point-wise confidence band, and its boundaries

Figure 2.6 shows the schematic of the 99% point-wise confidence band, in which 2.6a shows the comparison between the nonparametric density family and the 99% point-wise confidence band; 2.6b includes the shapes of the true density and the optimal density estimation. By constructing the 99% point-wise confidence band, we aim to include the shape of the true density in the band. Nevertheless, since this is a

simulation-based element, with small datasets, the 99% confidence level is in fact under weak assumptions. Through multiple experiments, the author found that when $B \geq 10,000$, the confidence bands for the same small dataset from different simulation runs are almost identical.

Overall, the 99% point-wise confidence band have two major usages:

- **Provides inference for multimodal distributions**: when the dataset size is extremely small (like $n < 50$), multimodality in the result can be caused by the quality of sample itself and therefore is not enough to make the inference. When the dataset size is larger, the confidence band is of good value for the multimodal inference.

- **Helps generating the idea of the "Maximum Variance Density"**: this part is described in detail in the next section.

A good question here can be: *why a 99% confidence band, not an 100% one?* This is because in processes like bootstrap, they can do very good job in estimating confidence intervals, but not the extreme values.

2.7 The "Maximum Variance Density": Another Option

The "Maximum Variance Density" is an attempt in this thesis to better integrate the uncertainty of small dataset in the process of uncertainty propagation. The objective of propagating the maximum variance densities is to provide another reference for making safer decisions.

The ideas of introducing the maximum variance density came from a basic uncertainty characterization process. As can be seen from Figure 2.7, in the uncertainty characterization and propagation process, when we have absolutely no knowledge about the underlying distribution of a variable, with only a range from a to b, the use of *uniform distribution* gives a distribution with the largest variance. In that case,

we can say that uniform distribution contains the "maximum uncertainty" among all the distributions in the range of a to b. To further illustrate the relationship between variance and uncertainty, the right part of the Figure 2.18 contains two normal distributions with the same mean and different variances. Under such a condition, it can be observed that $\sigma_A^2 < \sigma_B^2$, and we can say that distribution B contains more uncertainty than distribution A.

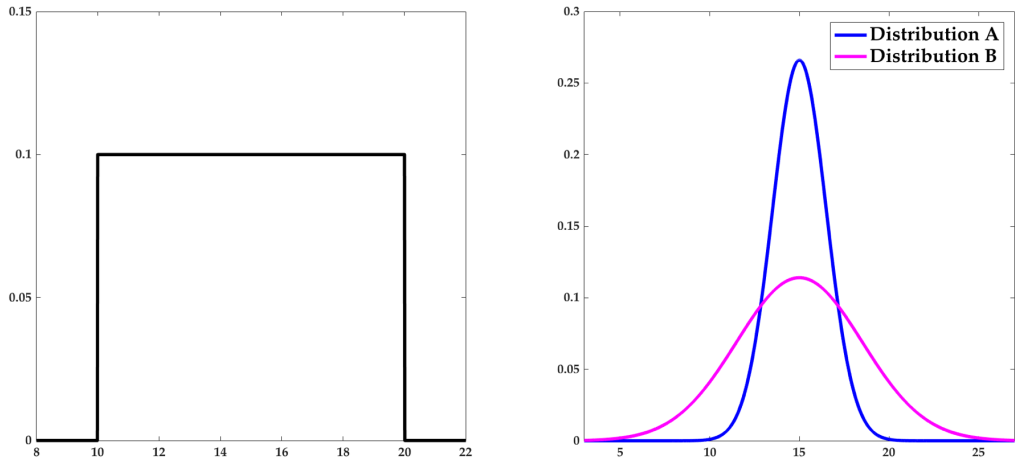


Figure 2.7: A uniform distribution (left) and two distributions with the same mean and different variances (right)

In our case, after using the steps mentioned in the previous section, we actually have some knowledge regarding the true density, in the form of the 99% confidence band. We can proceed with the assumption that the shape of the true density is included in the confidence band. So apart from the optimal density estimation, the author aims to provide another option (a sampling density) that contains the "maximum uncertainty" within the confidence band and represents the uncertainty in small dataset cases. The relationship between the 99% point-wise confidence band and the maximum variance density is shown in Figure 2.8. By propagating the maximum uncertainty densities for the uncertainty inputs, more potential extreme values in outputs can be caught, and the output distributions with larger variance are expected. In this way, the output distributions with larger variance can help us make

safer decisions for characteristics like the probability of success (POS).

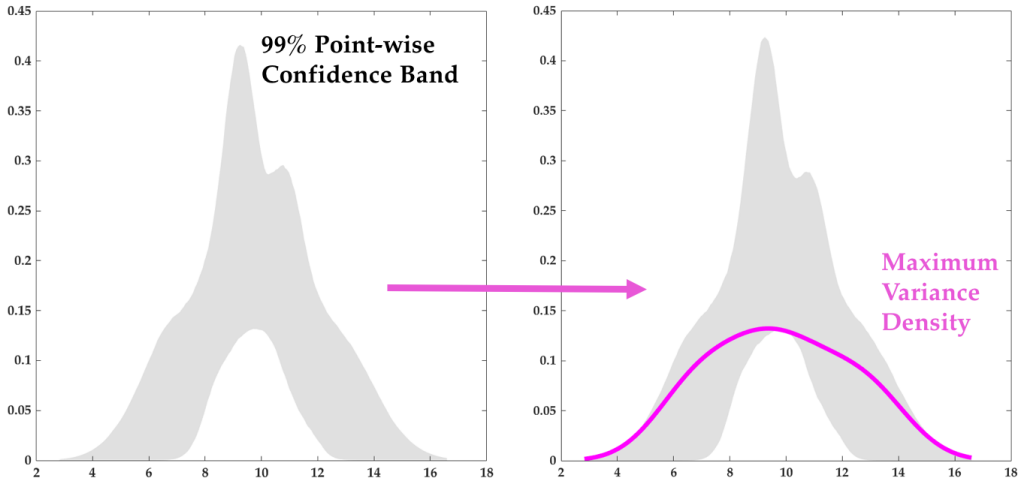


Figure 2.8: The relationship of the 99% point-wise confidence band and the maximum variance density

With the idea of finding such a density, now the real question becomes:

"How can we arrive at a maximum variance density from the confidence band?"

Initially, this maximum variance density is planned to be constructed through an optimization process set up, in which we

$$\begin{aligned} &\text{Maximize} \quad \sigma_q^2 \\ &\text{Subject to} \quad \int_\Omega q(x)dx - 1 = 0 \\ &\qquad\qquad L(x) < q(x) < U(x), \forall\ x \in \Omega \end{aligned} \tag{2.23}$$

where σ_q^2 is the variance of the target density, $L(x)$ and $U(x)$ are the lower and upper bounds of the confidence band.

Later it was discovered that identifying this maximum variance density from such an optimization set up is not the best way, for the following two reasons:

1. From the perspective of statistics, it doesn't make much sense to identify a density from such a band, and it will make more sense to obtain one from the simulation (bootstrap) process.

2. Even if a density with the maximum variance is obtained from an optimization process, the result may not best represent some characteristics of the true density, such as mean and skewness.

Finally, the maximum variance density is obtained from re-integrating information from the nonparametric density family. The complete process of obtaining the maximum variance density can be seen in Figure 2.9. Basically, after getting the whole nonparametric density family (with 10,000 densities), we select the densities whose variance is within the top 1% largest of the whole family. After that, an average nonparametric density of all these selected densities is obtained and used as the maximum variance density.

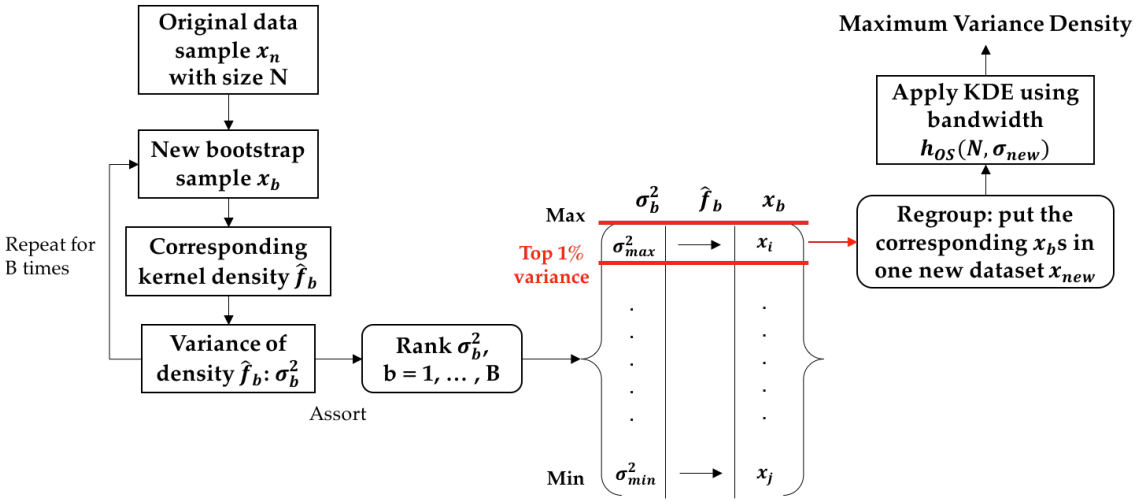


Figure 2.9: The process of obtaining the maximum variance density from the nonparametric family

In the process of getting the average nonparametric density of the top 1% largest variance densities, we add the corresponding bootstrap datasets of those selected densities to form a large dataset x_{new} with a total of $NB/100$ data, and apply the KDE with bandwidth $h_{OS}(N, \sigma_{new})$ to get the maximum variance density in the one-dimensional formulation.

2.8 Comparison Between the Maximum Variance and the Student's t-Distributions

As introduced before, the use of Student's t-Distribution on small datasets is not always appropriate, because it assumes that the observations are normally distributed. Nevertheless, as for the attempt to include uncertainty when modeling small datasets, the maximum variance density in this work do share some similarities with the Student's t-Distribution.

The Student's t-Distribution was derived by William Gossett, who published the work under the pseudonym 'Student' in 1908 [35]. It is used to describe how members of a *small sample* are distributed when selected randomly from a normal distribution. There are an infinite number of Student's t-Distribution, one for each degree of freedom ν, specified by [35]:

$$p(t,\nu) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\Gamma(\nu/2)}\left[1+\frac{t^2}{\nu}\right]^{-(\nu+1)/2} \tag{2.24}$$

where $\nu = n-1$ denotes the degrees of freedom and Γ is the gamma function [36]. An example of the Student's t-Distribution is shown in Figure 2.10.

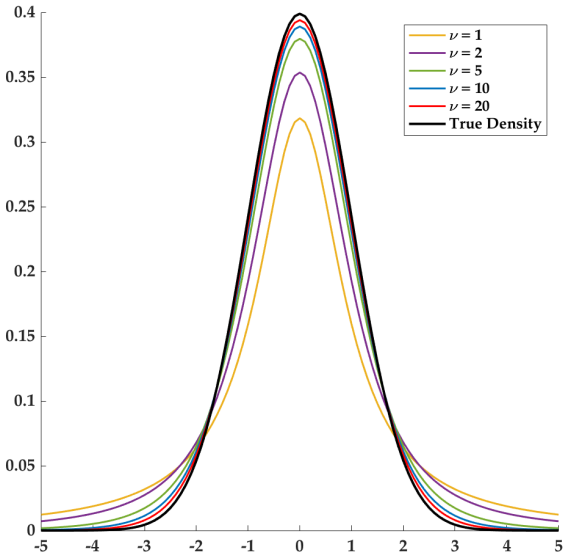


Figure 2.10: Student's t-Distribution with different degrees of freedom

The Student's t-Distribution is used for small samples (say, $n < 30$). As the degrees of freedom (also sample size) increase, the t-distribution approaches the normal distribution [36]. As can be observed from Figure 2.10, for small ν (or n), the t-distribution underestimates the normal distribution in the center and overestimates it on the tails [35].

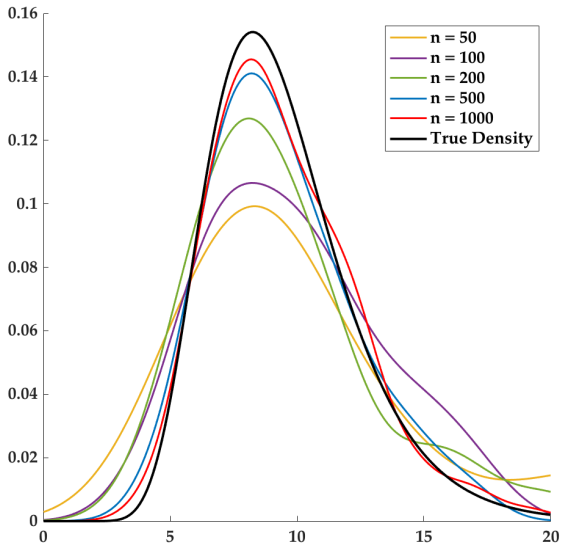


Figure 2.11: Some maximum variance densities of the same distribution with different sample sizes

On the other hand, a plot of some maximum variance densities of the same Log-normal Distribution with different sample sizes is shown in Figure 2.11. It can be observed that for such a skewed distribution, as the sample size n increases, the maximum variance densities approach the true density. They also tend to underestimates the center of the true density and overestimates it on the tails.

Similar characteristics of the Student's t-Distribution and the maximum variance density display similar thinkings in dealing with the uncertainty tied with small samples. Both of them show the degree of confidence that the majority of sample should fall within a certain range. Densities with heavier tails (or larger variance in our case) are prone to producing more values that are far from the mean, and therefore contain more uncertainty.

2.9 Formulation of the 1-D Solution

With all the elements described in the previous sections, final formulation of the one-dimensional solution consists of two solutions: optimal sampling density and maximum variance density. Figure 2.12 provides a complete picture of of all the elements in the result of the one-dimensional formulation, and the shape of the true density.

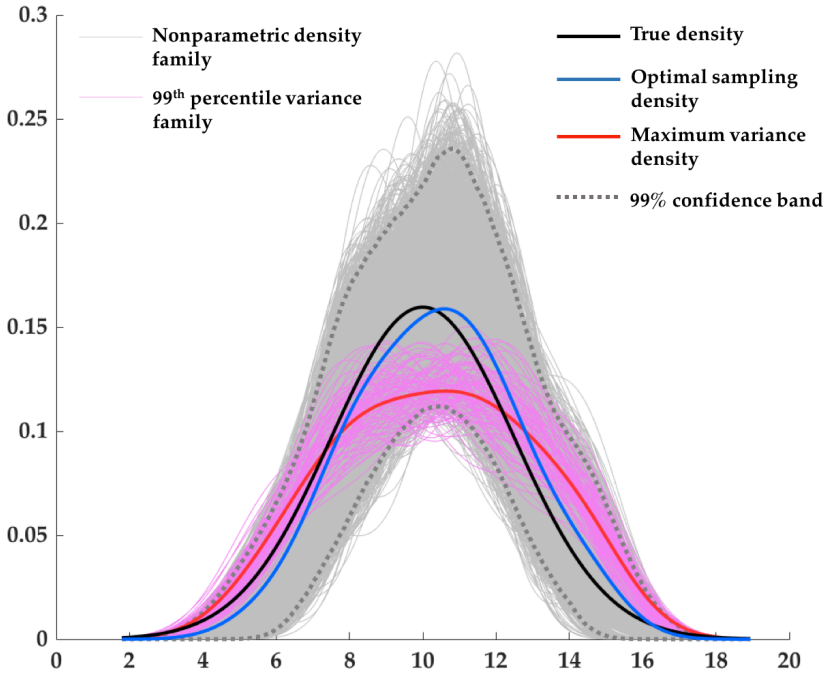


Figure 2.12: Main elements in the nonparametric one-dimensional estimation

There are five major elements in the one-dimensional result:

1. **Optimal Sampling Density**: calculated through KDE and various bandwidth selection methods (rules of thumb and cross-validation). This density is used for creating samples that represents the best estimation for the true density.

2. **Maximum Variance Density**: calculated through bootstrap and iterative KDE processes. It is used for creating sample that represents a density with

the maximum variance believed based on original data.

3. **Nonparametric Density Family**: calculated through bootstrap and KDEs. This density family is used to construct the confidence band, and act as the whole population to select the densities with top 1% largest variance.

4. **99% Percentile Variance Family**: selected through the whole nonparametric density family. This density family is used to generate the maximum variance density through re-grouping the corresponding bootstrap datasets and KDE.

5. **99% Confidence Band**: calculated through the whole nonparametric density family from bootstrap and KDE. The 99% point-wise confidence band can be used to make inference on the shape and multimodal property of the true density.

In the meantime, all the processes in the one-dimensional formulation are summarized and shown in the flow diagram in Figure 2.13. Key components in each of the density formulations are summarize below:

- **Optimal Sampling Density**: the most crucial component here is the selection of bandwidth for KDE towards the optimal density. The range of bandwidth h for cross-validation is formed by a lower bound (h_{SROT}) and an upper bound (h_{OS}). Also, two types of data partition methods are selected as candidates and are used under difference dataset sizes. Details of this part can be found in Table 2.1.

- **Maximum Variance Density**: the starting point of this thread is the saved information (mainly bootstrap datasets and the variance of their corresponding kernel densities) of the new nonparametric density family. From those information, a new dataset consists of the summation of $B/100$ bootstrap datasets

is formed, and the maximum variance density is estimated based on the new dataset. Details of this part can be found in Figure 2.9.

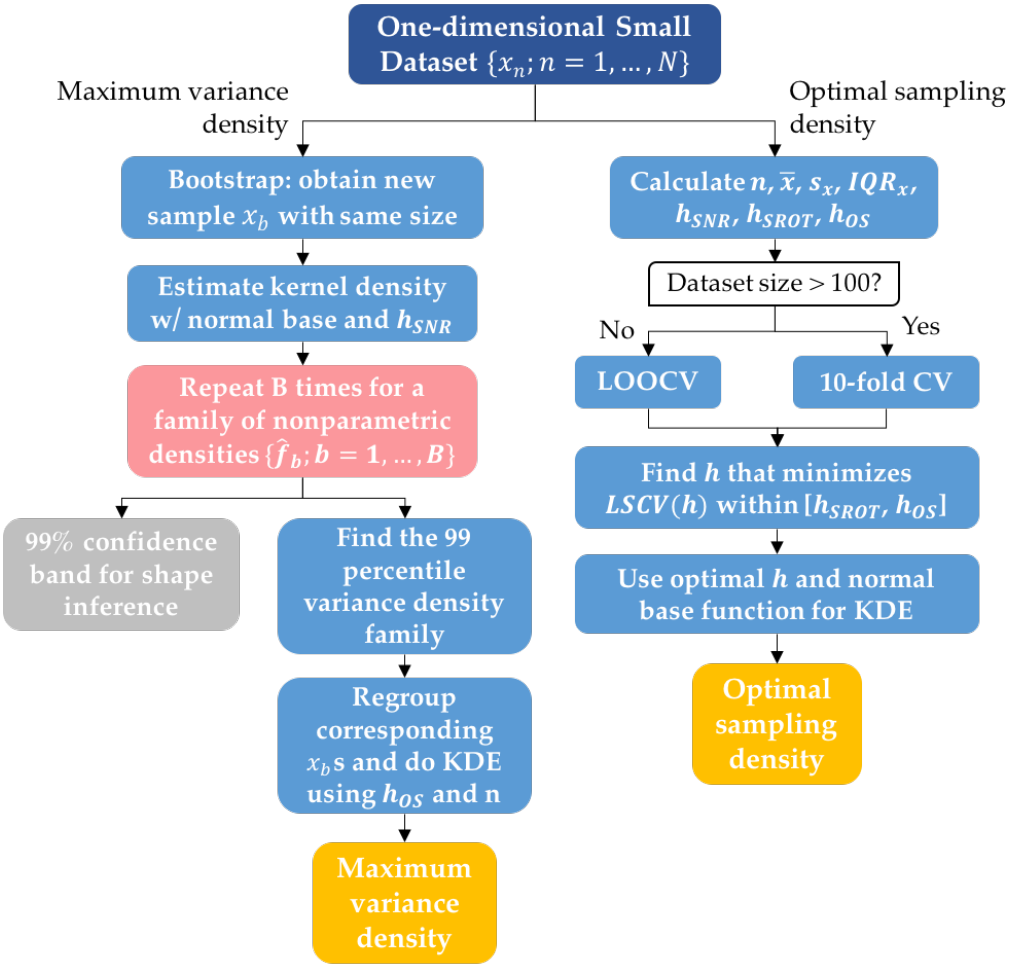


Figure 2.13: Flow diagram of the complete one-dimensional density estimations process

Based on the flow diagram in Figure 2.13, the algorithm for the one-dimensional estimations is presented in Algorithm 1. This algorithm is a simplified one only for uncertainty propagation: it takes the original dataset as the input, and exports the two one-dimensional densities.

Algorithm 1 One-dimensional Nonparametric Density Estimations

Input: A sample of size n: X_1, X_2, . . . , X_n
Output: Two sampling densities: optimal density $\hat{f}_{opt}(x)$, and maximum variance density $\hat{f}_{mvar}(x)$

Get basic characteristics of the dataset: $\bar{x}$, s_x, n, IQR_x.
Set the number of Bootstrap runs to be B.
Calculate the Rules of Thumb bandwidths for the dataset: h_{SNR}, h_{SROT}, h_{OS}
Set the cross validation boundary: $LB = h_{SROT}$, $UB = h_{OS}$
for $LB \leq h \leq UB$ **do**
    if $n \leq 100$ **then**
        Use Leave-one-out cross validation (LOOCV) partition
    else
        Use 10-fold cross validation (10-Fold CV) partition
    end if
    $LSCV(h) = \int (\hat{f}_h(x))^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)$
end for
$h_{opt} \leftarrow h$ that minimize $LSCV(h)$
$\hat{f}_{opt}(x) = \frac{1}{nh_{opt}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_{opt}}\right)$
for $i = 1$ to B **do**
    Generate Bootstrap sample $B[i]$ with the same size n, calculate $s[i]$ and save
    $h_{SNR}[i] \leftarrow h_{SNR}(n, s[i])$
    $\hat{f}_B[i] = \frac{1}{nh_{SNR}[i]} \sum_{j=1}^{n} K\left(\frac{x - B[i]_i}{h_{SNR}[i]}\right)$
    $\sigma[i] = Var(\hat{f}_B[i])$
end for
$P_{99} \leftarrow$ the $99th$ percentile of σ
Create a new empty array MVA
for $j = 1$ to B **do**
    if $\sigma[j] \geq P_{99}$ **then**
        Add data: $MVA \leftarrow MVA + B[j]$
    end if
end for
Get the variance and size of MVA: s_{MVA}
$h_{mvar} \leftarrow h_{OS}(n, s_{MVA})$
$\hat{f}_{mvar}(x) = \frac{1}{nh_{mvar}} \sum_{i=1}^{n_{MVA}} K\left(\frac{x - MVA_i}{h_{lvar}}\right)$

2.10 Test Cases for the 1-D Formulation

A total of four test cases were designed and executed to examine the effectiveness of the one-dimensional formulation. Overall, the one-dimensional formulation can be deemed effective if the following criteria are satisfied:

1. For the **optimal sampling density**: its two major statistical central moments (mean and standard deviation) should be close to the counterparts of the true density. If the true density is skewed, since skewness is a more rigorous measurement, the optimal sampling density should at least have a skewness with the correct sign (positive or negative).

2. For the **maximum variance density**: its first central moment (mean) should be close to the mean of the true density. In the meantime, its standard deviation must be greater than the standard deviation of the optimal sampling density. If the true density is skewed, the maximum variance density should also at least have a skewness with the correct sign (positive or negative).

3. For the **99% Confidence Band**: it should include the shape of the true density. This part is mainly examined visually.

4. For **multimodality**: the one-dimensional formulation should have the capability of estimating multimodality for some $n > 0$, depending on the shape of the actual true density.

Four distributions with different characteristics are selected as the true densities. The reasons for their selection are stated below:

- **Normal Distribution** $X \sim \mathcal{N}(10, 1.5^2)$: the most common probability distribution; symmetric (Skewness = 0)

- **Lognormal Distribution** $X \sim Lognormal(2.20, 0.30^2)$: has a positive skew compared to the normal distribution

- **Rayleigh Distribution** $X \sim Rayleigh(5)$: also positively skewed. This Rayleigh distribution has a sharp change on its shape near $x = 0$, which makes it even more difficult to estimate some details

- **Nonparametric Distribution**: a nonparametric distribution with two modes. It will test the formulation's capability of simulating multimodality

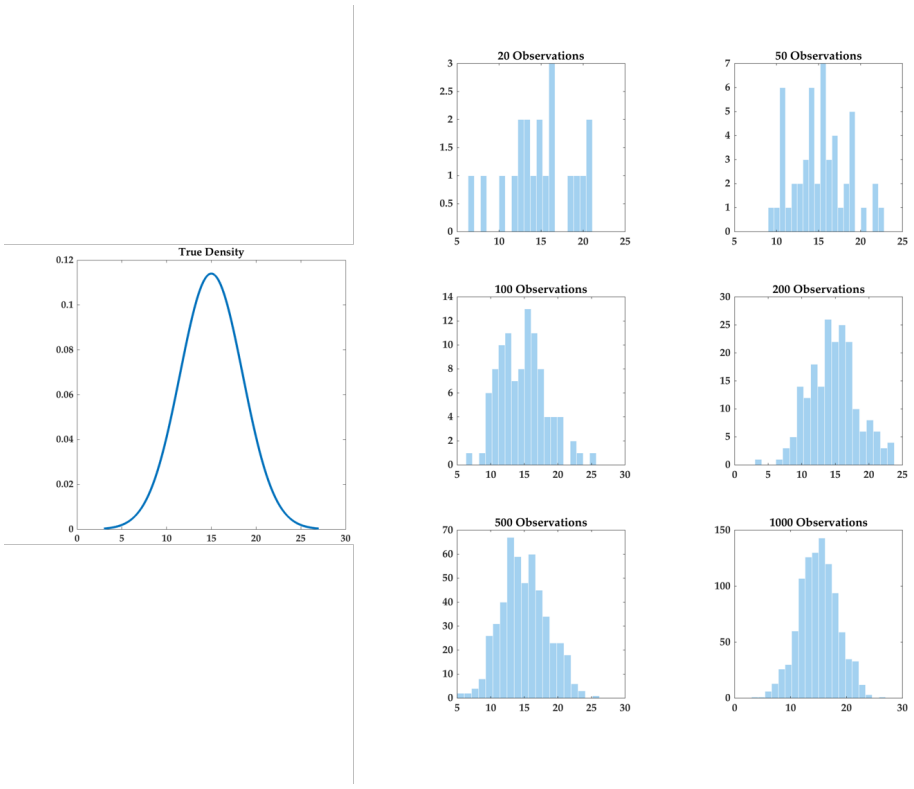


Figure 2.14: Random samples from the same density with different numbers of observations

For the testing process, we take random samples from the true density with the number of observations n =20, 50, 100, 200, 500, 1000, respectively. We then apply the one-dimensional formulation to estimate the true density from the samples, and compare the results visually and via different statistical moments.

2.10.1 Test Case 1: Normal Distribution $X \sim \mathcal{N}(10, 1.5^2)$

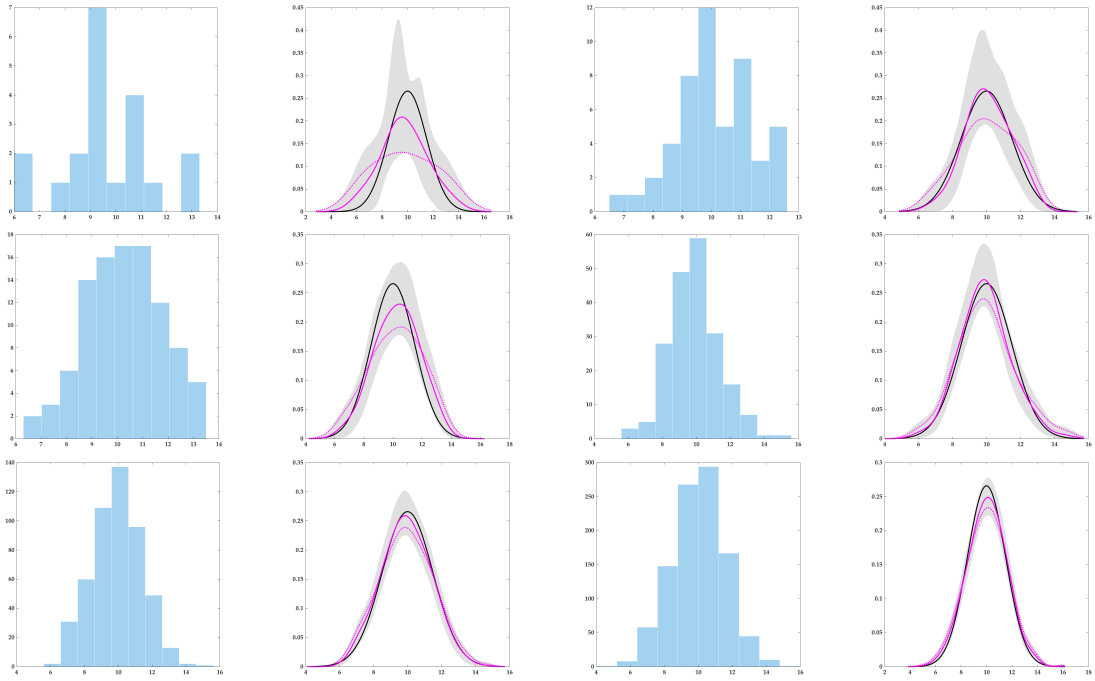


Figure 2.15: Test Case 1: Normal Distribution $X \sim \mathcal{N}(10, 1.5^2)$ for 20 observations (top left), 50 observations (top right), 100 observations (middle left), 200 observations (middle right), 500 observations (bottom left) and 1000 observations (bottom right)

Table 2.2: Test Case 1: Normal Distribution $X \sim \mathcal{N}(10, 1.5^2)$: comparison of the moments among the original density, optimal sampling density, and maximum variance sampling density

20 Data	**Ori.**	**Opt.**	**Max Var.**	**200 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	10	9.739	9.791	***Mean***	10	9.928	9.971
Std.	1.5	2.001	2.692	***Std.***	1.5	1.585	1.811
Skew.	0	0.0456	0.0265	***Skew.***	0	0.217	0.243
50 Data	**Ori.**	**Opt.**	**Max Var.**	**500 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	10	10.107	10.037	***Mean***	10	9.929	9.932
Std.	1.5	1.471	1.809	***Std.***	1.5	1.532	1.656
Skew.	0	-0.037	-0.146	***Skew.***	0	0.049	0.071
100 Data	**Ori.**	**Opt.**	**Max Var.**	**1000 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	10	10.262	10.209	***Mean***	10	10.025	10.013
Std.	1.5	1.649	1.926	***Std.***	1.5	1.611	1.711
Skew.	0	-0.121	-0.147	***Skew.***	0	-0.026	0.017

2.10.2 Test Case 2: Lognormal Distribution $X \sim Lognormal(2.20, 0.30^2)$

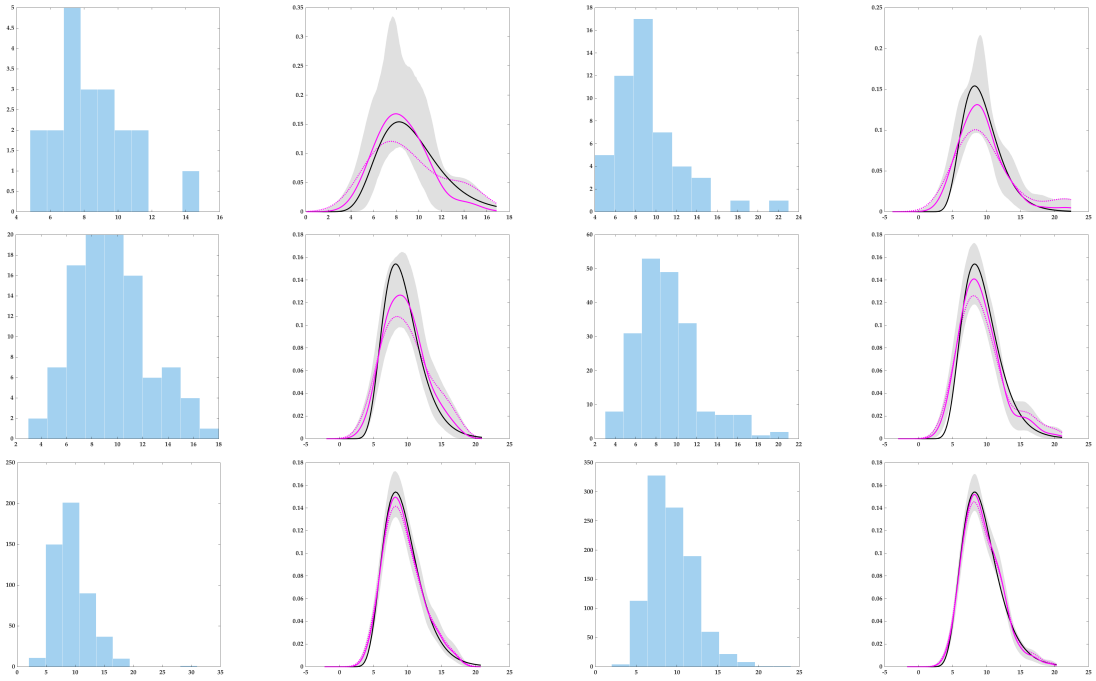


Figure 2.16: Test Case 2: Lognormal Distribution $X \sim Lognormal(2.20, 0.30^2)$ for 20 observations (top left), 50 observations (top right), 100 observations (middle left), 200 observations (middle right), 500 observations (bottom left) and 1000 observations (bottom right)

Table 2.3: Test Case 2: Lognormal Distribution $X \sim Lognormal(2.20, 0.30^2)$: comparison of the moments among the original density, optimal sampling density, and maximum variance sampling density

20 Data	**Ori.**	**Opt.**	**Max Var.**	**200 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	9.44	8.499	9.096	***Mean***	9.44	9.112	9.418
Std.	2.89	2.439	3.481	***Std.***	2.89	3.252	3.789
Skew.	0.95	0.448	0.361	***Skew.***	0.95	0.839	0.844
50 Data	**Ori.**	**Opt.**	**Max Var.**	**500 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	9.44	9.298	10.182	***Mean***	9.44	9.376	9.537
Std.	2.89	3.674	5.202	***Std.***	2.89	3.037	3.585
Skew.	0.95	1.105	0.962	***Skew.***	0.95	1.073	1.895
100 Data	**Ori.**	**Opt.**	**Max Var.**	**1000 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	9.44	9.534	9.799	***Mean***	9.44	9.432	9.499
Std.	2.89	3.133	3.691	***Std.***	2.89	2.889	3.108
Skew.	0.95	0.387	0.335	***Skew.***	0.95	0.773	0.891

2.10.3 Test Case 3: Rayleigh Distribution $X \sim Rayleigh(5)$

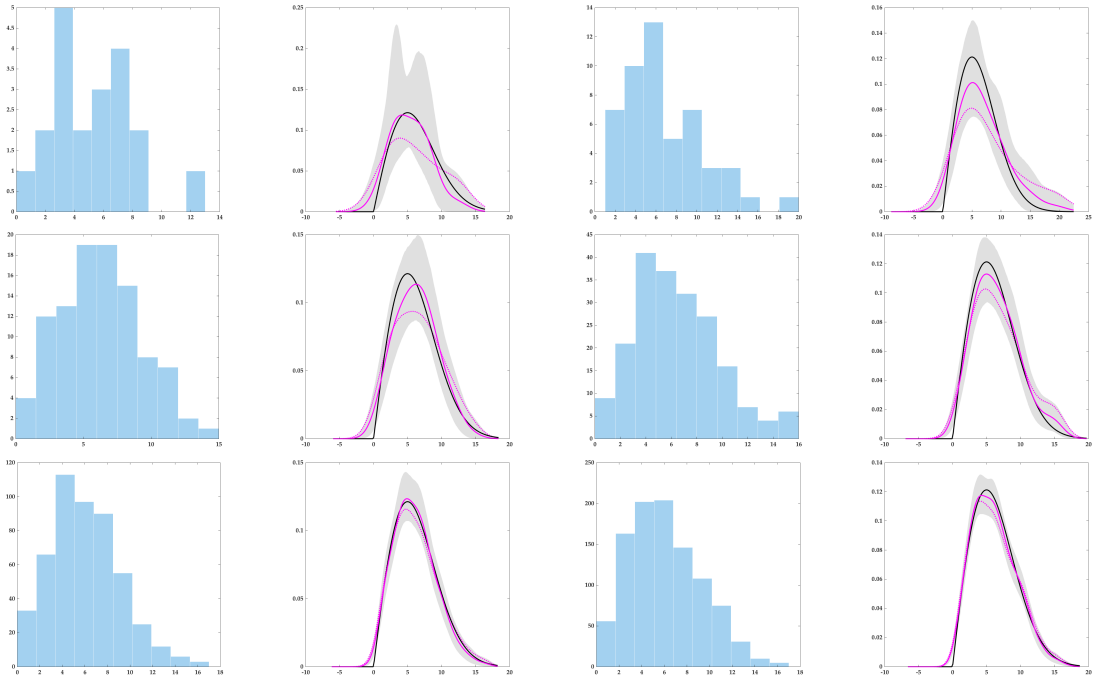


Figure 2.17: Test Case 3: Rayleigh Distribution $X \sim Rayleigh(5)$ for 20 observations (top left), 50 observations (top right), 100 observations (middle left), 200 observations (middle right), 500 observations (bottom left) and 1000 observations (bottom right)

Table 2.4: Test Case 3: Rayleigh Distribution $X \sim Rayleigh(5)$: comparison of the moments among the original density, optimal sampling density, and maximum variance sampling density

20 Data	**Ori.**	**Opt.**	**Max Var.**	**200 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	6.26	5.426	6.033	***Mean***	6.26	6.449	6.721
Std.	3.27	3.181	4.369	***Std.***	3.27	3.582	4.046
Skew.	0.62	0.321	0.276	***Skew.***	0.62	0.498	0.504
50 Data	**Ori.**	**Opt.**	**Max Var.**	**500 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	6.26	6.791	7.669	***Mean***	6.26	6.067	6.199
Std.	3.27	4.387	5.699	***Std.***	3.27	3.208	3.496
Skew.	0.62	0.709	0.649	***Skew.***	0.62	0.499	0.551
100 Data	**Ori.**	**Opt.**	**Max Var.**	**1000 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	6.26	6.247	6.439	***Mean***	6.26	6.109	6.181
Std.	3.27	3.348	3.872	***Std.***	3.27	3.302	3.497
Skew.	0.62	0.206	0.241	***Skew.***	0.62	0.477	0.508

2.10.4 Test Case 4: Nonparametric Distribution

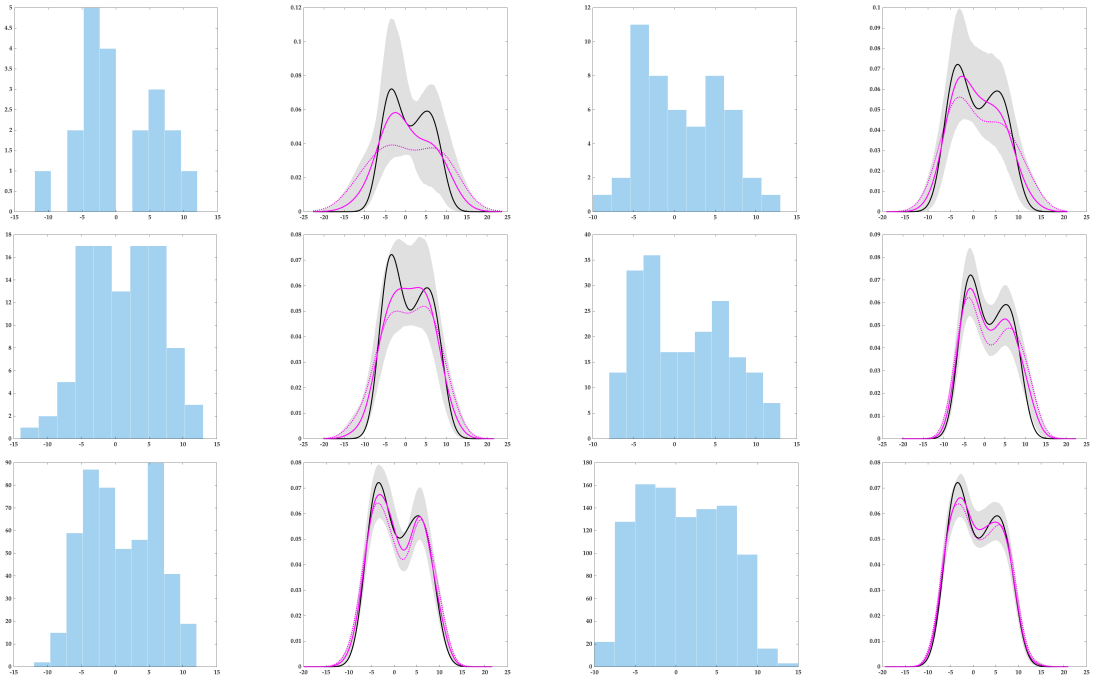


Figure 2.18: Test Case 4: Nonparametric Distribution for 20 observations (top left), 50 observations (top right), 100 observations (middle left), 200 observations (middle right), 500 observations (bottom left) and 1000 observations (bottom right)

Table 2.5: Test Case 4: Nonparametric Distribution: comparison of the moments among the original density, optimal sampling density, and maximum variance sampling density

20 Data	**Ori.**	**Opt.**	**Max Var.**	**200 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	0.88	0.487	0.554	***Mean***	0.88	1.147	1.424
Std.	5.19	6.732	8.618	***Std.***	5.19	5.738	6.256
Skew.	0.11	0.098	-0.044	***Skew.***	0.11	0.185	0.152
50 Data	**Ori.**	**Opt.**	**Max Var.**	**500 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	0.88	0.886	1.291	***Mean***	0.88	0.849	0.894
Std.	5.19	5.617	6.666	***Std.***	5.19	5.493	5.779
Skew.	0.11	0.204	0.199	***Skew.***	0.11	0.086	0.069
100 Data	**Ori.**	**Opt.**	**Max Var.**	**1000 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean	0.88	0.838	0.829	***Mean***	0.88	0.839	0.915
Std.	5.19	5.713	6.617	***Std.***	5.19	5.284	5.499
Skew.	0.11	-0.029	-0.091	***Skew.***	0.11	0.115	0.099

2.11 Results from the Test Cases

In general, results from the four 1-D test cases show that the 1-D formulation is effective as the criteria settled for the optimal sampling density, the maximum variance density and the 99% confidence band are basically satisfied. Although there are some randomness in small samples, elements in the 1-D formulation can generally give close and reasonable estimations. Below we summarize the observations made from the four test cases:

1. **Test Case 1**: with samples from a normal distribution, the **optimal sampling densities** have close estimations in all of mean, standard deviation and skewness. The **maximum variance sampling densities** can also provide good estimations on mean and skewness, while give reasonably larger standard deviations than the optimal sampling densities. As the sample size grows, the maximum variance sampling densities are getting closer to the optimal sampling densities as expected. The **99% confidence bands** can do very good job in including the true density. The confidence band shrinks as the sample size grows, which is also expected.

2. **Test Case 2**: with samples from a skewed Lognormal distribution, apart from good estimations in mean and standard deviation, the **optimal sampling densities** and **maximum variance sampling densities** demonstrate good capabilities in estimating skewness. Even if skewness is a more rigorous measurement and is very sensitive to small differences, the two sampling densities can give good estimations while keeping 100% correct in the sign.

3. **Test Case 3**: the Rayleigh distribution creates more difficulties in estimating the details of its shape, because it has a sharp change (almost perpendicular to x-axis) at $x = 0$. As shown in the test case results, this indeed creates

difficulties for the three 1-D elements to estimate the shape around this corner. Nevertheless, estimations of the three moments are still in an excellent quality.

4. **Test Case 4**: this test case is important as it is out of the capability range for parametric-based methods without using mixture models. All of the three 1-D elements examined perform well, and with an enough amount of data (say, $n > 200$), all the three elements show good capability of estimating multimodality.

2.12 Comparisons of the Formulations: Parametric vs. Nonparametric

To conclude the one-dimensional chapter, a comparison of the parametric-based and nonparametric-based formulations are shown in Figure 2.19.

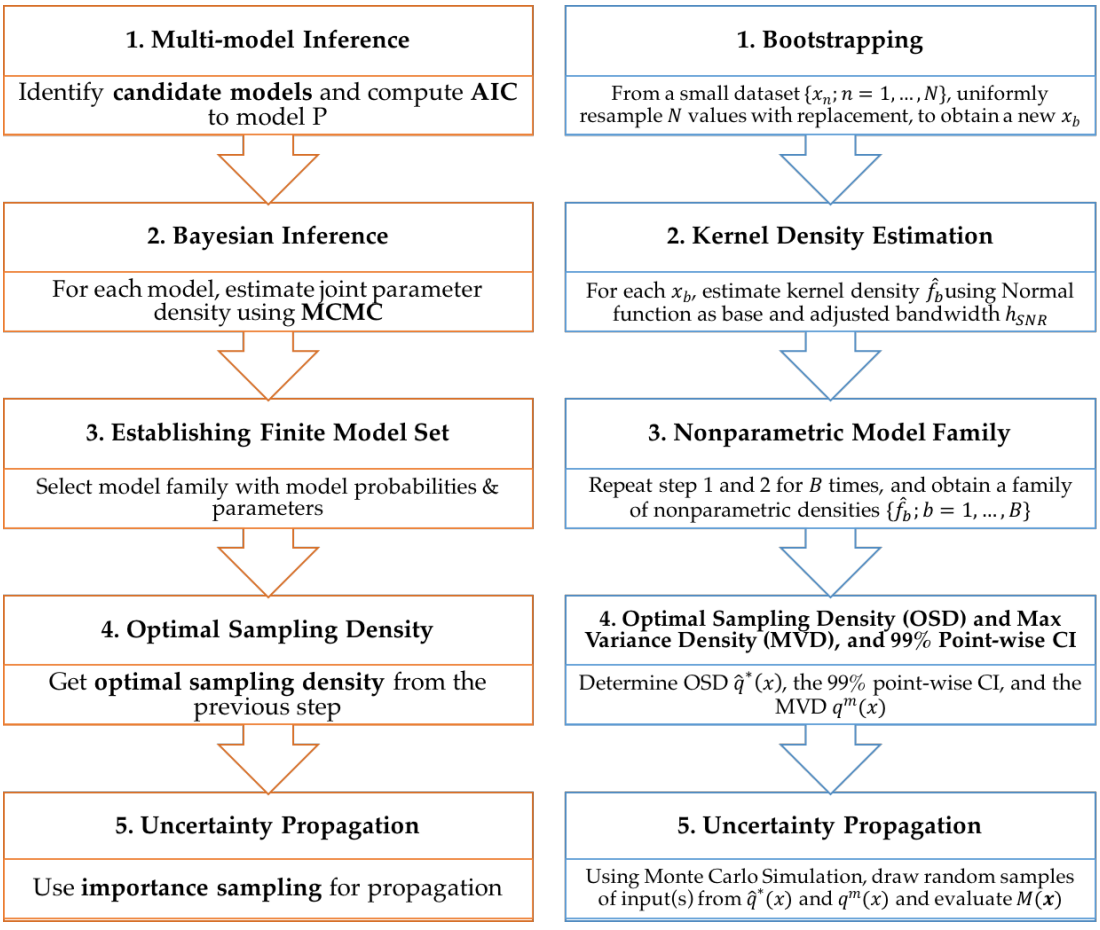


Figure 2.19: Comparison of the parametric-based (left) and nonparametric-based (right) one-dimensional formulations

CHAPTER 3
2-D NONPARAMETRIC DENSITY ESTIMATION

3.1 Motivation for 2-D Modeling

Even though there's an effective formulation for density estimation and uncertainty propagation for small datasets for the one-dimensional case, that may still be not enough to give good estimations of the output distributions. And the reason is simple: among the n inputs in a systems model, *correlations may exist* among some of them.

When propagating and analyzing uncertainty associated with a complex systems model in industries like aerospace engineering, the high degree of complexity in inputs and the model itself can make some common assumptions during uncertainty analysis, such as the independence of random variables, no longer valid. In those cases, ignoring the correlation among related random variables may have a substantial effect on probabilistic assessment and uncertainty quantification results [10].

3.1.1 Example 1: Parametric Uncertainty Assessment for AEDT

As introduced before, the Aviation Environmental Design Tool (AEDT) is a tool for assessing fleet wide fuel burn, emissions, and noise impacts. Since uncertainties exist in input parameters of AEDT, a project was conducted to propagate those uncertainties through the model, and quantify uncertainties in AEDT outputs [37]. The research team performed correlation analysis on the key AEDT input parameters, and identified that some correlations exist among the input parameters.

Result of the analysis showed some correlations as follows:

- **TOGW vs. Thrust**: strong positive correlation
- **Overall Pressure Ratio (OPR) vs. NOx**: strong positive correlation

- **Bypass Ratio (BPR) vs. NOx**: weak negative correlation

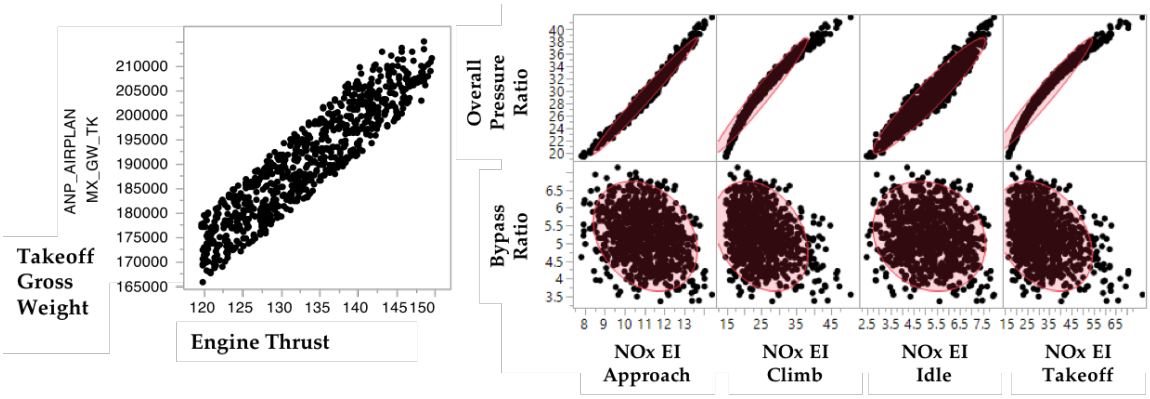


Figure 3.1: Correlations among the AEDT input parameters: TOGW vs. Thrust, OPR vs. NOx, and BPR vs. NOx

Details of the correlations are shown in the scatter plots in Figure 3.1. Realized that output uncertainty can be overestimated or underestimated if those correlations are ignored in the uncertainty analysis, researchers of the project utilized copulas theory to model the dependencies and generated joint probability distributions.

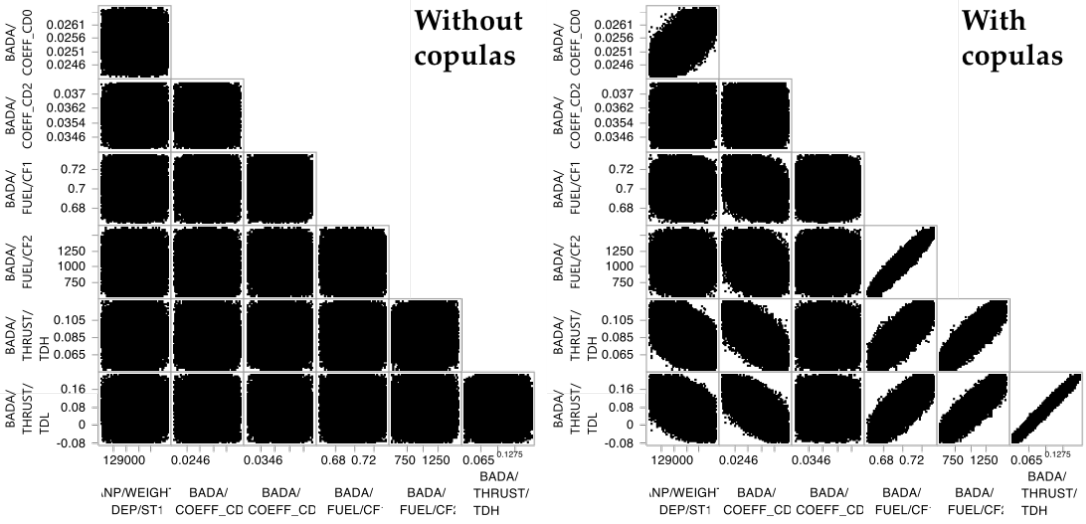


Figure 3.2: Comparison of the Monte Carlo Samples, with copulas (right), and without copulas (left)

Figure 3.2 contains an comparison between the Monte Carlo samples for the input parameters, with and without copulas. It can be observed that copulas captured some

strong to moderate positive correlations and weak negative correlations among the input parameters, and the use of copulas makes a big difference in the simulation process (normal distributions were used to model individual input parameters, and Gaussian copulas were used to model correlations).

When such a large difference exist between the Monte Carlo samples, it can be imagined that output distributions from the two groups of simulations are also different. Take the *Departure NOx* and *Approach NOx* (two of the outputs) as an example: it can be observed from Figure 3.3 that when copulas are used in the process, output distribution with less uncertainty was obtained. This result coincides with the common sense in this case, and illustrates the importance of two-dimensional modeling for inputs which can model correlations.

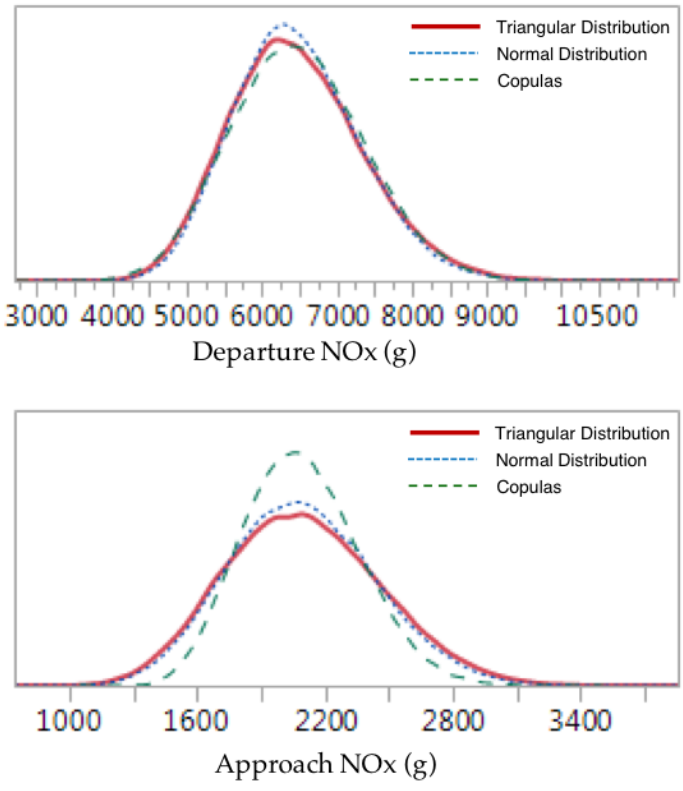


Figure 3.3: Output distributions for terminal NOx using different modeling methods

3.1.2 Example 2: Evaluation of Technology Impacts on System Performance

A study conducted by Turab Zaidi in 2014 [10] also showed the significance of change in output uncertainty when copulas are used in the modeling the simulation process. When uncertainty dependence structures between some technology impact factors were modeled and simulated, their impact on predicted aircraft performance is different from the modeling and simulation when technology impact factors are assumed independent.

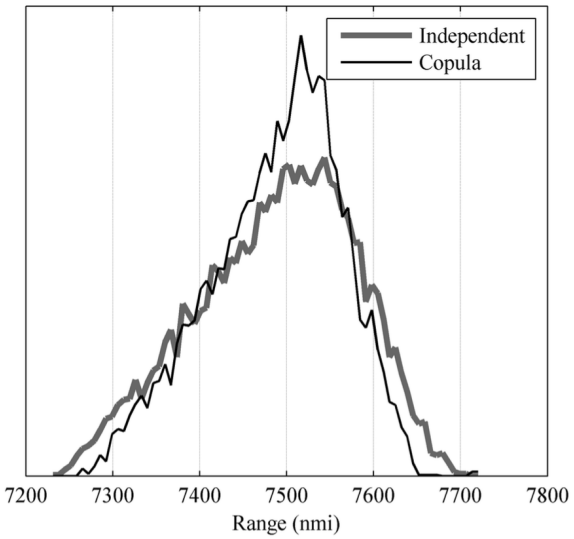


Figure 3.4: Empirical PDF of range subject to independent and Frank copula correlated technology impact factors [10]

As can be seen from Figure 3.4, the two simulations reveal obvious difference in the variance of output distributions, in which the distribution resulting from independent random variable assumptions having greater uncertainty than its copula-deployed counterpart.

The two examples demonstrate the necessity of modeling correlations in such uncertainty propagation processes. In the end, what we need here is a two-dimensional nonparametric formulation to capture correlations among the inputs, and propagate uncertainty under the influence of small datasets.

3.2 Correlation Coefficient and Correlation Test for Small Dataset

3.2.1 Correlation and Correlation Coefficient

Correlation

In the area of science, the term *correlation* is used to refer to an association, connection, or any form of relationship, link or correspondence [38]. Two types of common correlation include *linear correlation* and *nonlinear correlation*. Correlation is said to be linear if the ratio of change between x and y is constant, and it happens when all the points on a scatter diagram tend to lie near a straight line. Nonlinear (or curvilinear) correlation describes the situation when the ratio of change between x and y is not constant. In the nonlinear correlation case, all the points on a scatter diagram tend to lie near a smooth curve [39] [40].

In statistics, however, correlation is used to describe the extent of linear association between two continuous random variables. It is measured by a statistic called *correlation coefficient*, which represents the strength of linear association between the variables in question [38].

Correlation Coefficient

The *correlation coefficient* is a dimensionless quantity that lies between -1 and 1. The Pearson correlation coefficient (PCC) is a product moment coefficient that is used in this work. A value of PCC close to 1 indicates a strong positive linear relationship, and a value close to -1 indicates a strong negative linear relationship. A value close to 0 indicates no linear relationship between the variables [39]. A PCC value of -1 or +1 indicates a perfect linear relationship. Some examples of datasets and their PCC values can be found in Figure 3.5. The PCC for population is defined as

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X-u_X)(Y-u_Y)]}{\sigma_X \sigma_Y} \tag{3.1}$$

where u_X and u_Y are expected values and σ_X and σ_Y are standard deviations; $\text{cov}(X, Y)$ is the covariance of X and Y.

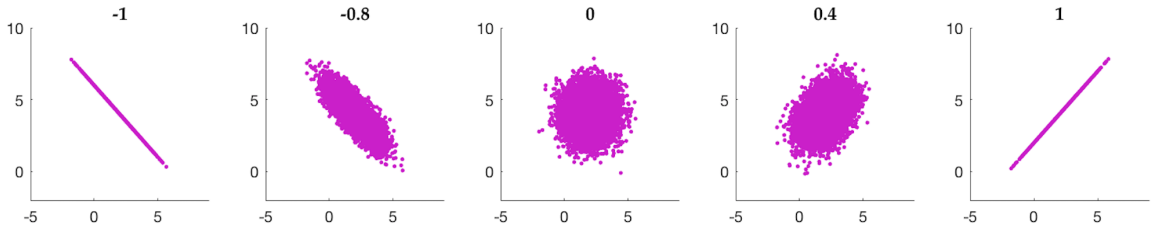


Figure 3.5: Examples of datasets and their correlation coefficients: (from left to right) $r = -1$, $r = -0.8$, $r = 0$, $r = 0.4$, $r = 1$

For an actual sample that has n pairs of data ($[x_1, y_1]$, $[x_2, y_2]$, ... , $[x_n, y_n]$), the sample PCC is then given by [39]

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of u_X and u_Y.

One limitation for the PCC is that correlation includes a variety of relationships between X and Y, while as a correlation coefficient, the PCC can only represent the degree of linear relationship.

3.2.2 Correlation Test for Small Dataset

Since the PCC actually has its limitation, and we are facing the small dataset case in this work, the correlation test must be carefully designed such that the best method can be applied to certain situations.

In general, there are three situations as of the correlation between two variables X and Y: *no correlation*, *linear correlation*, and *curvilinear (non-linear) correlation*. For each situation, the corresponding solutions are:

- **No correlation**: apply 1-D density estimations on X and Y independently

- **Linear correlation**: apply 1-D estimations on marginal densities of X and Y, and copula to capture the correlation and form the joint density
- **Curvilinear correlation**: apply the 2-D multivariate kernel density estimation

For our small dataset case, lack of information can create additional difficulties in judging the correlation status between X and Y. As a result, much more tests other than just a simple correlation test must be done to draw the conclusion. In order to better classify the data into the three correlation situations stated above, two test are used here, in which each test consists of several statistical components.

Test 1: is there a curvilinear correlation between X and Y?

The first test is to examine if there is a curvilinear correlation between X and Y. This is because if the correlation between X and Y is dominated by linear correlation, then all the rest of the tasks here will be used on the strength of linearity. If X and Y are curvilinearly related, copula will not be used.

The best way to distinguish between curvilinear correlation and linear correlation is *curvilinear regression* [41]. In this regression process, we add different powers of the X variable (X, X^2, X^3) to an equation to see whether the Coefficient of Determination R^2 increase significantly. We first do a linear regression and fit the equation $\hat{Y} = a + b_1 X$ to the data. After that, we do quadratic ($\hat{Y} = a + b_1 X + b_2 X^2$) and cubic ($\hat{Y} = a + b_1 X + b_2 X^2 + b_3 X^3$) regressions and test the increase in R^2. We limit the degree of polynomial no larger than 3 because in most of the cases, that is enough for us to judge a curvilinear correlation. It is normal that as we add in X^2 and X^3, R^2 will increase because of the higher order term. But the key question here is whether R^2 will increase significantly compared with a linear regression. In the end, we can say that the curvilinear correlation is more dominant if

$$\max\{R^2_{quad}, R^2_{cub}\} > R^2_{line} + 0.25 \tag{3.3}$$

Figure 3.6 is an example of curvilinear correlation. With the same dataset and different regressions, it can be observed that in this case, the cubic correlation is more significant than linear and quadratic correlations (R^2 for cubic regression is 0.6324, a lot larger than around 0.36 for linear and quadratic regressions). In a situation like this, correlation between X and Y will be classified as curvilinear correlation.

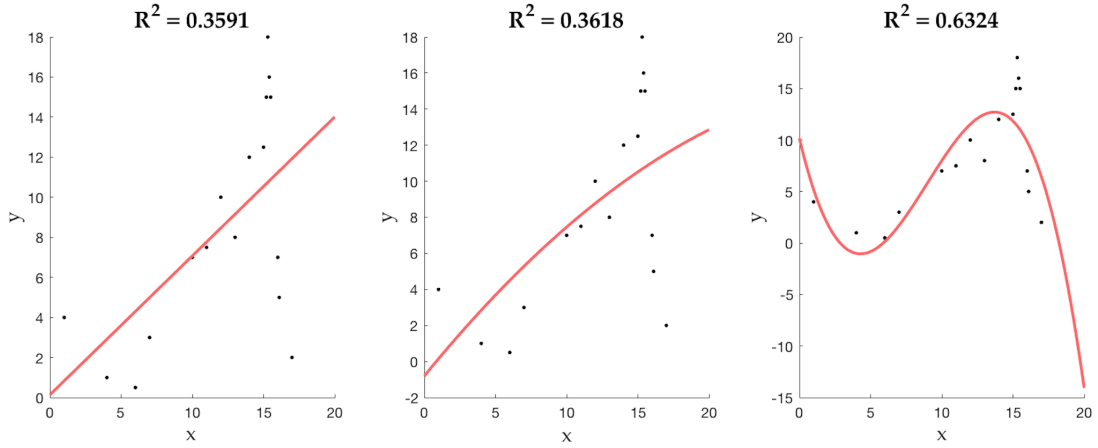


Figure 3.6: Example: result of a dataset with linear, quadratic and cubic regression, and their coefficient of determination (R^2)

Test 2: is there a strong enough linear correlation between X and Y?

If results from test 1 is not enough to conclude a curvilinear correlation, a second test is needed to discover if the linear correlation is strong enough. As highlighted in previous sections, uncertainty always exists in small dataset case. Therefore, an additional criterion is needed to statistically infer the strength of linear correlation. Confidence intervals can give us additional information that indicate the strength.

A combination of two statistics (two types of confidence intervals) is used to make the linear correlation decision:

- *Bootstrap confidence interval*: using the original dataset, 1,000 new bootstrap samples (B_1, B_2, ... , B_{1000}) are created. For each new bootstrap dataset, the correlation coefficient PCC is computed and saved. The 95% confidence interval

of PCC is then computed through the final distribution of 1,000 PCCs.

- *Confidence interval through transformation* [39]: a 95% confidence interval can be calculated using Fisher's z transformation:

$$z_r = \log_e \frac{1+r}{1-r} \tag{3.4}$$

$$[\text{Lowerbound}' \quad \text{Upperbound}'] = [z_r - 1.96 \cdot \frac{1}{\sqrt{n-3}}, \quad z_r + 1.96 \cdot \frac{1}{\sqrt{n-3}}] \tag{3.5}$$

$$[\text{Lowerbound} \quad \text{Upperbound}] = [\frac{e^{2\cdot\text{Lowerbound}'} - 1}{e^{2\cdot\text{Lowerbound}'} + 1}, \quad \frac{e^{2\cdot\text{Upperbound}'} - 1}{e^{2\cdot\text{Upperbound}'} + 1}] \tag{3.6}$$

Both the two confidence intervals are used in the inference. Two statistics in between are used and compared to make a final decision: center of the confidence interval through transformation (t_1), and median of bootstrap confidence interval (t_2). When the minimum of two statistics ($\min\{|t_1|, |t_2|\}$) is greater than a certain value that stands for negligible correlation, X and Y are linearly correlated.

Table 3.1: Interpretation of correlation coefficient within $-1 < r < 1$

Correlation Coefficient	**Interpretation**
$0.9 < \|r\| < 1.0$	very high positive/negative correlation
$0.7 < \|r\| \leq 0.9$	high positive/negative correlation
$0.5 < \|r\| \leq 0.7$	moderate positive/negative correlation
$0.3 < \|r\| \leq 0.5$	low positive/negative correlation
$0.0 < \|r\| \leq 0.3$	negligible correlation

Table 3.1 [38] contains an interpretation of correlation coefficient that is widely accepted by the community of statistics. It can be seen that when the absolute value of PCC is smaller than 0.3, the correlation can be judged as negligible correlation. At the end of this test 2, we can say that the linear correlation is strong enough when

$$\min\{|t_1|, |t_2|\} > 0.3 \tag{3.7}$$

A complete correlation test process that contains test 1 and test 2 is assorted and

shown in Figure 3.7. The role of the correlation test in the two-dimensional solution can be found in Figure 3.8.

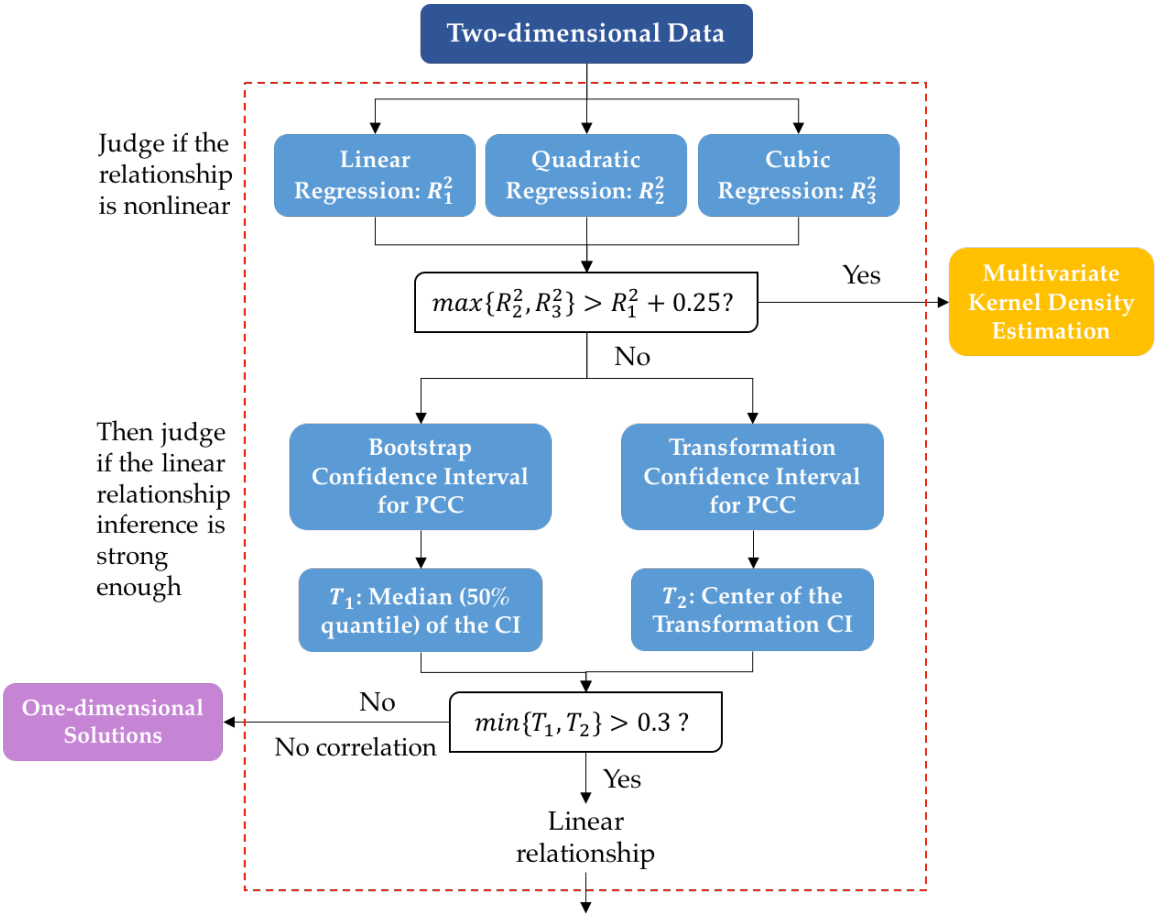


Figure 3.7: Complete correlation test process, including test 1 (for nonlinearity) and test 2 (for the strength of linearity)

Based on the flow diagram in Figure 3.7, the algorithm for a complete correlation test under the influence of small datasets is presented in Algorithm 2. This algorithm takes the original two-dimensional dataset as the input, and exports one of the three correlation cases:

1. **Case 1**: *curvilinear relationship* between x and y
2. **Case 2**: *no relationship* between x and y
3. **Case 3**: *linear relationship* between x and y

Algorithm 2 Two-dimensional Correlation Judgment Based on Small Dataset

Input: A two-dimensional dataset (x_i, y_i), $i = 1, ..., n$
Output: One of the three cases:
Case 1 - curvilinear relationship between x and y
Case 2 - no relationship between x and y
Case 3 - linear relationship between x and y

Judge curvilinear regressions: apply linear, quadratic, and cubic regressions
Get coefficients of determination: R^2_{line}, R^2_{quad}, and R^2_{cub}
for $i = 1$ to B **do**
    Generate Bootstrap sample $B[i]$ with the same size n, get $\bar{x}_i$ and $\bar{y}_i$
    Get PCC for $B[i]$: $PCC[i] = \rho_{X,Y} = \dfrac{\sum_{j=1}^{n}(x_j - \bar{x}_i)(y_j - \bar{y}_i)}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x}_i)^2 \sum_{j=1}^{n}(y_j - \bar{y}_i)^2}}$
end for
$T_1 = \text{median}(PCC)$
Get a default PCC for data: $r_d = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$
Fisher's z transformation: $z_{rd} = \log_e = \dfrac{1 + r_d}{1 - r_d}$
$[\text{LB}' \ \ \text{UB}'] = [z_{rd} - 1.96 \cdot \dfrac{1}{\sqrt{n-3}}, \ \ z_{rd} + 1.96 \cdot \dfrac{1}{\sqrt{n-3}}]$
$[\text{LB} \ \ \text{UB}] = [\dfrac{e^{2 \cdot \text{LB}'} - 1}{e^{2 \cdot \text{LB}'} + 1}, \ \ \dfrac{e^{2 \cdot \text{UB}'} - 1}{e^{2 \cdot \text{UB}'} + 1}]$
$T_2 = 0.5 \cdot (LB + UB)$
if $\max\{R^2_{quad}, R^2_{cub}\} - 0.25 > R^2_{line}$ **then**
    Case = Case 1
else if $\min\{|T_1|, |T_2|\} > 0.3$ **then**
    Case = Case 3
else
    Case = Case 2
end if

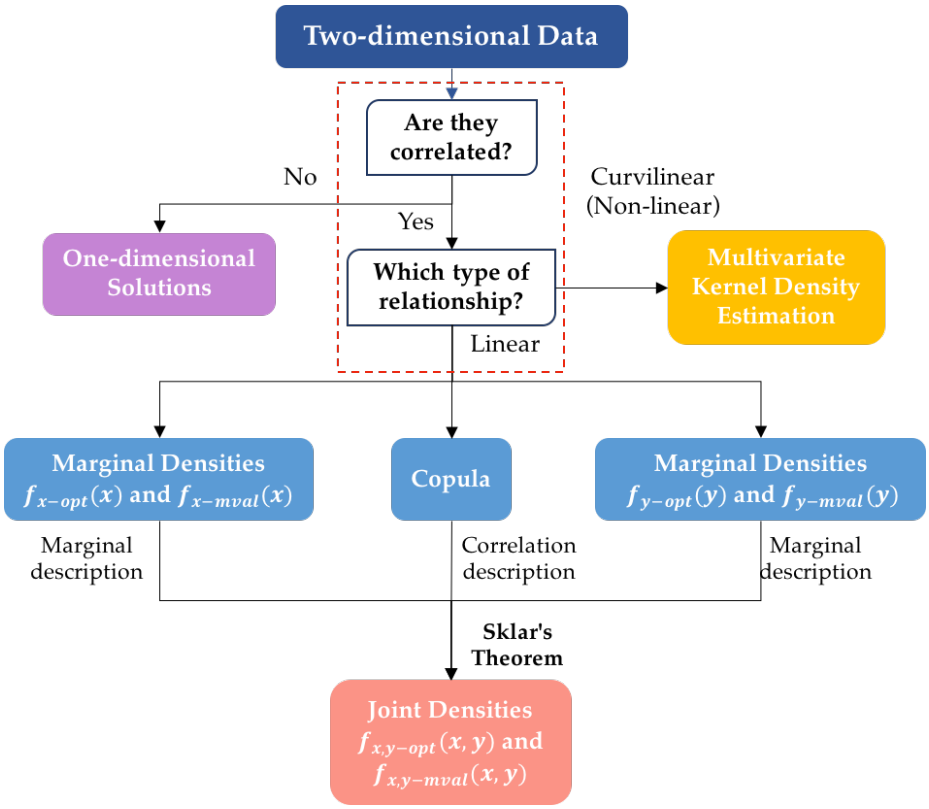


Figure 3.8: The role of the correlation test in the complete two-dimensional solution process

3.3 Multivariate Kernel Density Estimation (For 2-D)

3.3.1 Nonparametric Multivariate Density Estimation

Nonparametric multivariate density estimation techniques do not require assumption about the data, and has two major types: the *popular kernel method* which uses locally tuned radial basis (e.g., Gaussian) functions to interpolate the multidimensional density, and the *exploratory projection pursuit method* which interprets the multidimensional density through the construction of several 1-D densities along highly "interesting" projections of multidimensional data [42].

Given a set of multi-dimensional observed data, the task of multivariate density estimation is to find an function $\hat{f}$ that best approximates the true probability density function f. There are two constraints in this process: positivity ($\hat{f}(x) \geq 0$) and sum-

to one ($\int \hat{f}(x)dx = 1$) constraints [42].

An alternative to the kernel estimator is the mixture model, where the underlying density is assumed to have the form

$$g(x) = \sum_{i=1}^{k} p_i g_i(x; \theta_i) \tag{3.8}$$

where $\{p_i, i = 1, ..., k\}$ are weights for the multivariate normal g_i ($\theta_i = \{u_i, \Sigma)i\}$) and are constrained so that $p_i > 0$ and $\sum_{i-1}^{k} p_i = 1$. The use of mixture models are often motivated by heterogeneity or the presence of distinct subpopulations in observed data [43].

3.3.2 Kernel Density Estimation for Bivariate Data

The multivariate kernel density estimation is an extension of the univariate KDE introduced in the last chapter. When we have a d-dimensional data set with sample size n as

$$X_i = (X_{i1}, ..., X_{id})^T, i = 1, ..., n \tag{3.9}$$

the goal is to estimate the density $\hat{f}$ of $X_i = (X_{i1}, ..., X_{id})^T$: $\hat{f}(\mathbf{x}) = \hat{f}(x_1, ..., x_d)$.

The kernel density estimator in d-dimensions is defined by

$$\begin{aligned} \hat{f}_h(\mathbf{x}) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K(\frac{\mathbf{x} - \mathbf{X}_i}{h}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K(\frac{x_1 - X_{i1}}{h}, ..., \frac{x_d - X_{id}}{h}) \end{aligned} \tag{3.10}$$

where K is a multivariate kernel function with d arguments, and h is the same for each components [44]. With the extension of bandwidth $h = (h_1, ..., h_d)^T$, we have

$$\hat{f}_h(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1 ... h_d} K(\frac{x_1 - X_{i1}}{h_1}, ..., \frac{x_d - X_{id}}{h_d}) \tag{3.11}$$

For two-dimensional data, $\mathbf{x} = (x_1, x_2)^T$, $\mathbf{X}_i = (X_{i1}, X_{i2})^T$, $i = 1, 2, ..., n$. There are many choices for kernel function K, such as Gaussian kernel, Spherical/radial-

symmetric kernel, Epanechnikov kernel, etc [44] [45]. The selection of K is not a key discussion point here.

The choice of bandwidth is much more crucial than the choice of kernel function. Compared with the bandwidth in one-dimensional KDE which is a number, bandwidth for a multivariate kernel density estimator is a matrix H (non-singular). In a simple rule-of-thumb case, we use a bandwidth matrix proportional to $\hat{\Sigma}^{-1/2}$, where $\hat{\Sigma}$ is the covariance matrix of the data. In cases like this, we use bandwidth matrices to adjust for correlation between the components [44]. Common bandwidth selection methods include plug-in method and cross-validation method which are similar to the one-dimensional case. More advanced bandwidth selection methods for multivariate KDE include Bayesian approaches like the Markov chain Monte Carlo (MCMC) [46].

In the two-dimensional case, the bandwidth matrix H is a 2×2 matrix, that needs to be positive definite and symmetric [45]

$$\mathbf{H} = [h_1^2 \quad 0; 0 \quad h_2^2] or \ \mathbf{H} = [h_1^2 \quad h_{12}; h_{12} \quad h_2^2] \tag{3.12}$$

In this thesis, the exploration on the two-dimensional estimation mainly focus on using copulas to capture the correlation between variables. Therefore, the use of multivariate kernel density estimation only falls in the situation when correlation case 1 (curvilinear relationship between x and y) happens.

An adaptive kernel density estimator based on linear diffusion processes called KDE via diffusion (Botev, 2010) is used in this work to visualize two-dimensional data and in the test processes [47]. This estimator uses adaptive smoothing by incorporating information from a pilot density estimate. A simulation example of this estimator can be found in Figure 3.9.

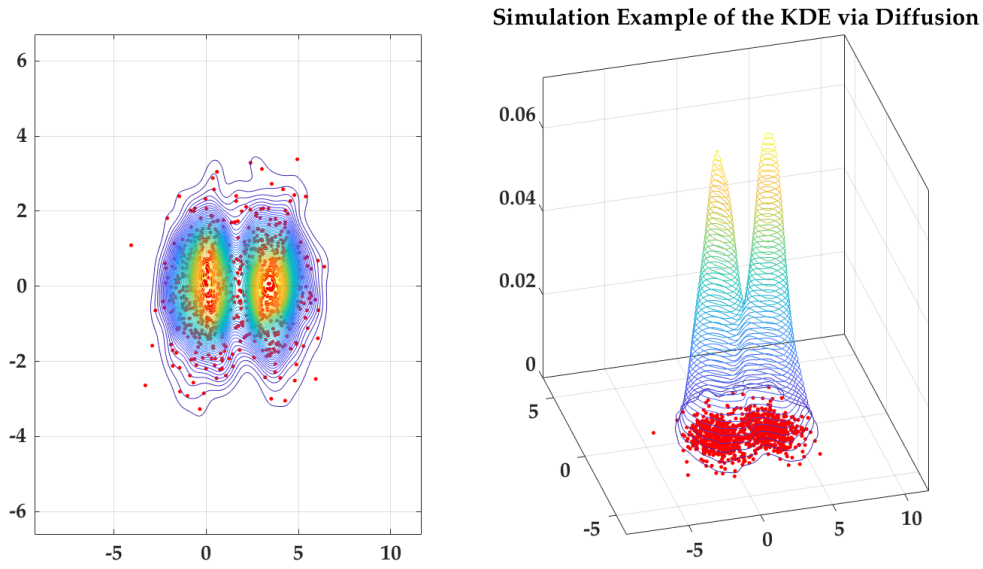


Figure 3.9: Simulation example of the KDE via diffusion estimator

3.4 Copulas

3.4.1 Copula Approach

There are a few approaches that can be used to estimate multivariate probability distributions: native multivariate distributions (Normal, Student-t, Wishart, Gamma, etc), mixture models, multivariate KDE, and marginal and conditional PDF [48]. Most of these approaches are estimations for joint densities directly, and have their own limitations. For example, the native multivariate distributions can not be used to model dependence structures for random vector with non-normal margins.

The *copulas (Latin: link)* methods model the dependence structure among variables in a different way, as they describe the joint distribution of X_1, X_2, ... , X_n by the marginal distributions $F_j(x)$ and a copula C [49]. Copulas have the power to describe the dependence explicitly, and links the marginal distributions together to form the joint distribution.

Definition 1. *Let* $X = (X_1, ..., X_n)$ *be a random vector with distribution function*

F and with marginal distribution functions F_i, $X_i \sim F_i$, $1 \leq i \leq n$. A distribution function C with uniform marginals on [0, 1] is called a "copula" of X if

$$F = C(F_1, ..., F_n) \tag{3.13}$$

Define C to be the distribution function of $(F_1(X_1), ..., F_n(X_n))$, since $F_i(X_i) \sim U(0,1)$ and C is a copula, we proceed to obtain [50]

$$\begin{aligned} C(u_1, ..., u_n) &= P(F_1(X_1) \leq u_1, ..., F_n(X_n) \leq u_n) \\ &= P(X_1 \leq F_1^{-1}(u_1), ..., X_n \leq F_n^{-1}(u_n)) \\ &= F_X(F_1^{-1}(u_1), ..., F_n^{-1}(u_n)) \end{aligned} \tag{3.14}$$

where F_i^{-1} denotes the generalized inverse of F_i, defined by

$$F_i^{-1}(t) = \inf\{x \in \mathrm{R}; F_i(x) \geq t\} \tag{3.15}$$

When $F_1(X_1)$, ... , $F_n(X_n)$ are independent

$$C(u_1, ..., u_n) = u_1 \times ... \times u_n \tag{3.16}$$

When $F_1(X_1)$, ... , $F_n(X_n)$ are completely dependent ($F_1(X_1) = ... = F_n(X_n)$ with probability 1)

$$C(u_1, ..., u_n) = \min\{u_1, ..., u_n\} \tag{3.17}$$

A complete process of the copula approach is shown in Figure 3.10. Since copula is defined as a function with uniform marginals on $[0, 1]$, a first transformation is needed such that marginal distributions are transformed to uniform distributions. With the uniform marginal distributions, copulas can be used to model dependency among the variables, and generate new samples from the modeled joint distribution. In the end, a transformation back to the original marginal distributions gives us data in the original scale.

There are three main classes of copulas: Archimedean, Gaussian, and t-copula [49]:

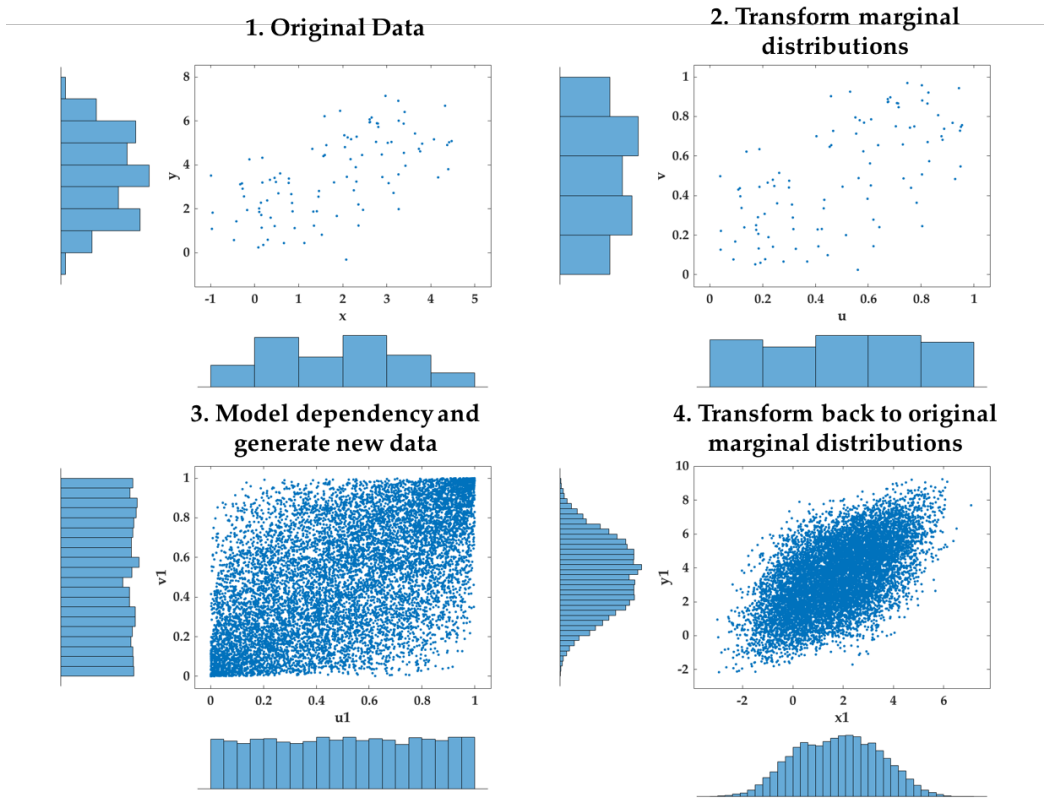


Figure 3.10: A complete four steps process of using copulas to model the dependency between two correlated variables

- **Archimedean copula**: has the form

$$C(u_1, ..., u_n) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2) + ... + \psi^{-1}(u_n)) \tag{3.18}$$

  for an appropriate generator function $\psi(\cdot)$. It includes the Clayton, Frank, Gumbel, and Joe copulas, each has a single parameter that controls the degree of dependence.

- **Gaussian copula**: is the copula when F_X in Equation 3.14 is a multivariate normal distribution $N_n(\mu, \Sigma)$

- ***t*-copula**: when F_X in Equation 3.14 is a multivariate t-distribution $t_v(\mu, \Sigma)$, C is a t-copula

The Archimedean class of copulas is by far the most popular in literature, for its simplicity in mathematical formulation and and ease to use [10]. Due to the

nonparametric nature of this thesis study, Gaussian copula and t-copula will not be used in modeling the dependency among variables. Through numerous experiments and tests, the Frank copula in the Archimedean class is used. It is worth mentioning that differences among the modeled dependence structures do exists when different copulas are used on the same dataset, as can be seen in Figure 3.11.

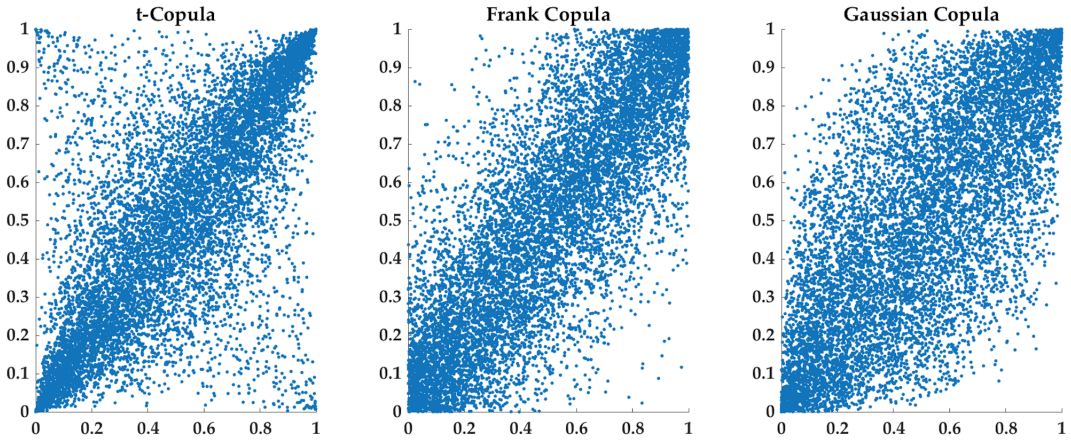


Figure 3.11: Dependency modeling on the same dataset with different types of copulas

3.4.2 Sklar's theorem

The copulas are important because of *Sklar's theorem* [49]. When C is a copula of F, then by definition of C [50]

$$\begin{aligned} F(x_1, ..., x_n) &= P(X_1 \leq x_1, ..., X_n \leq x_n) \\ &= P(F_1(X_1) \leq F_1(x_1), ..., F_n(X_n) \leq F_n(x_n)) \\ &= C(F_1(x_1), ..., F_n(x_n)) \end{aligned} \tag{3.19}$$

In the modeling process, Sklar's theorem allows us to separate the modeling of the marginal distributions $F_i(x)$ from the dependence structure C, because for any random variables $X_1, \ldots, X_n$, with joint cumulative distribution function (CDF)

$$F(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n) \tag{3.20}$$

and marginal CDFs

$$F_i(x) = P(X_i \leq x_i) \tag{3.21}$$

there exists a copula such that

$$F(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)) \tag{3.22}$$

And C is unique when each marginal distribution $F_i(x)$ is continuous [49].

When F and C are differentiable, Equation 3.22 satisfies

$$\frac{f(x_1, ..., x_n)}{f_1(x_1) \cdots f_n(x_n)} = c(F_1(x_1), ..., F_n(x_n)) \tag{3.23}$$

where c is the joint probability distribution function (PDF) of the copula

$$c(u_1, ..., u_n) = \frac{\partial^n}{\partial u_1 ... \partial u_n} C(u_1, ..., u_n) \tag{3.24}$$

Now with the Sklar's theorem, every joint PDF can be written as the product of n marginal PDFs and the PDF of the Copula. This theorem provides a great opportunity to extend the one-dimensional formulation of the work to a two-dimensional one which can capture correlation among the variables, because it allows us to model the marginal distributions separately, and form the joint distribution using a copula. Since this thesis work mainly deals with uncertainty associated with small datasets, a more statistically theoretical work uses different approach to apply Sklar's theorem in an imprecise setting [51]. They study the extension of Sklar's theorem when there is imprecision about the marginals by means of p-boxes and other elements.

3.5 The "Maximum Variance Density" in the 2-D Case

Like the maximum variance density in the one-dimensional formulation, the 2-D maximum variance density is also used to incorporate uncertainty which is inherent in small datasets. Compared to an optimal density estimation, the maximum variance density estimation in the 2-D case is also expected to cover a wider area and therefore

contains 'more uncertainty'.

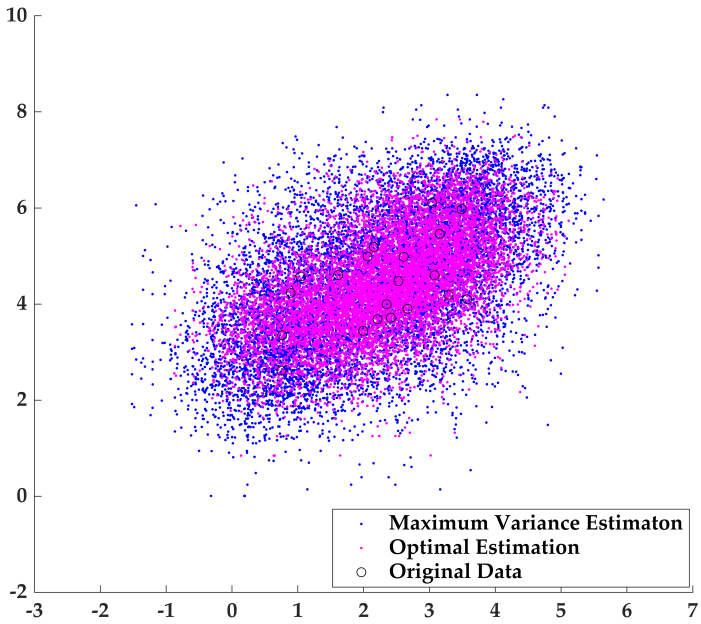


Figure 3.12: Example of optimal (magenta) and maximum variance (blue) density estimations on the same dataset: joint plot

As an example, a comparison of the optimal density estimation and the maximum variance density estimation on the same dataset is shown in Figure 3.12 (joint) and Figure 3.13 (separate). Apparently, as a density estimation that contains more uncertainty, the 2-D maximum variance density estimation covers a wider range in both x and y directions. When we sample from the maximum variance density and propagate uncertainty through a MCS-based process, it can be expected that the output distributions will have larger variances. Formations details of the two types of estimations are included in the next section.

3.6 Formulation of the 2-D Solution

Due to the use of copulas and Sklar's theorem, we are able to extend the 1-D formulation to a 2-D one which is capable of modeling dependencies. In this process we separate the modeling of marginal densities, and form the joint density using a copula. There are three main steps in the 2-D formulation (on data $\{(X_i, Y_i), i = 1, ..., n\}$):

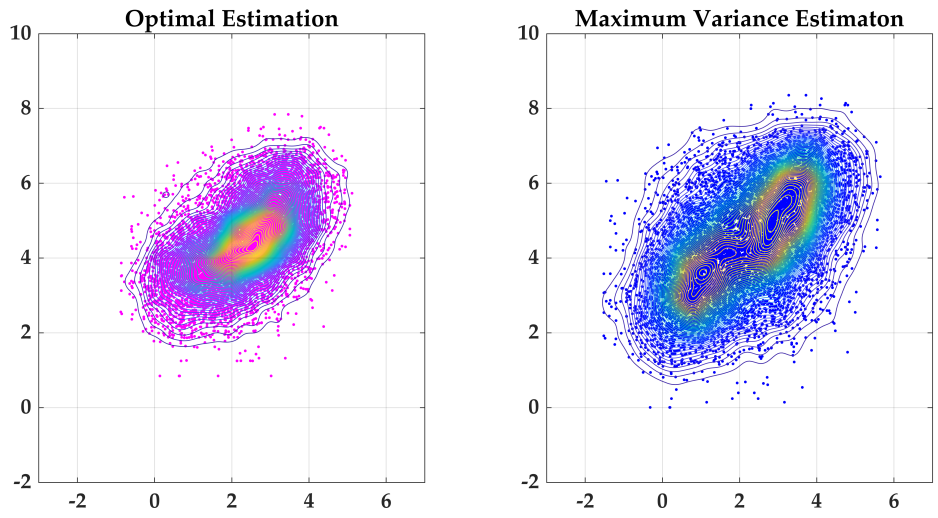


Figure 3.13: Example of optimal (magenta) and maximum variance (blue) density estimations on separate plots

1. **Model marginal probability densities**: the 1-D formulation will be used on the two dimensions separately. The 1-D optimal sampling densities and maximum variance densities are obtained as the marginal density estimations for x and y. This step one outputs four density estimations: $\hat{f}_{x_{opt}}(x)$, $\hat{f}_{y_{opt}}(y)$, $\hat{f}_{x_{mvar}}(x)$, $\hat{f}_{y_{mvar}}(y)$.

2. **Model dependency**: after a linear relationship between x and y is confirmed, the Frank copula is used to model the dependency between x and y. This step two outputs the type of copula (Frank) and the parameter that controls the degree of dependency.

3. **Form joint probability density**: Sklar's theorem is used to obtain the joint densities through the product of copula (C) and marginal densities. This last step outputs two joint densities - the joint optimal density ($\hat{f}_{x_{opt}}(x) \cdot \hat{f}_{y_{opt}}(y) \cdot C$), and the joint maximum variance density ($\hat{f}_{x_{mvar}}(x) \cdot \hat{f}_{y_{mvar}}(y) \cdot C$).

A flow diagram of the complete formulation of the 2-D solution can also be found in Figure 3.14.

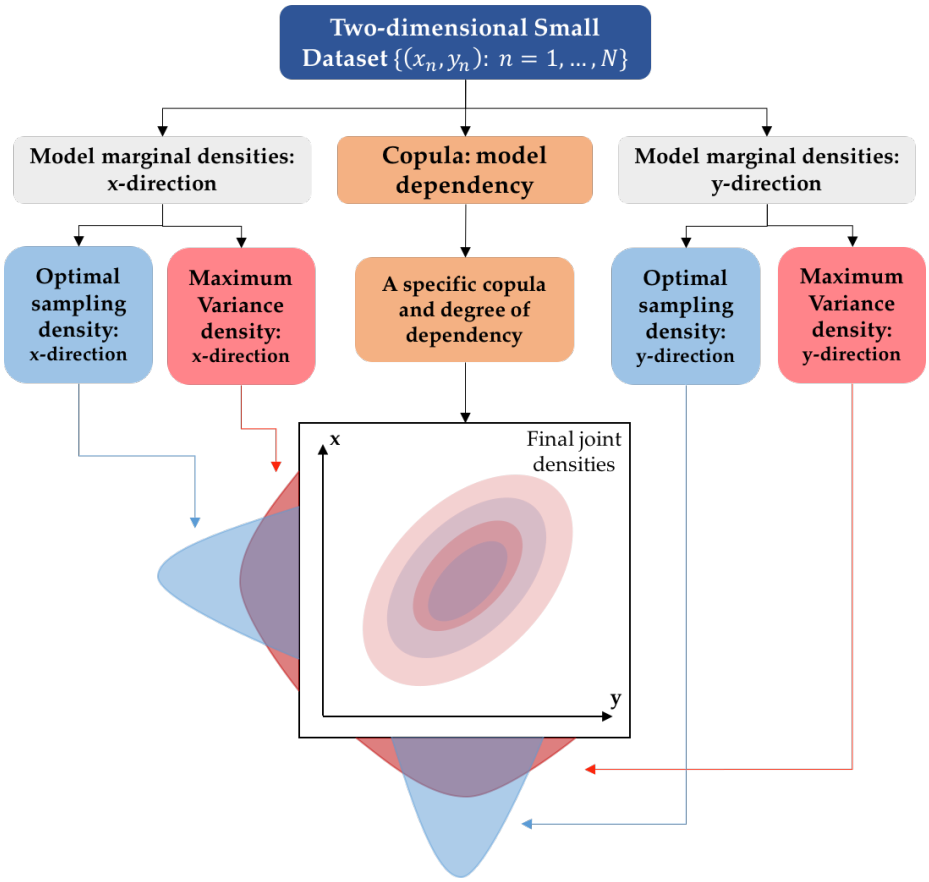


Figure 3.14: Flow diagram of the complete two-dimensional density estimations process (outputs are the two joint densities)

3.7 Test Cases for the 2-D Formulation

Like the test cases for 1-D formulation, the 2-D test cases are designed to examine the effectiveness of the two-dimensional formulation. In this process, we take random samples from the true 2-D density with the number of observations n =20, 50, 100, 200, 500, 1000 respectively, and then apply the two-dimensional formulation to estimate the true density from the samples. After that we compare the results visually and via different statistical moments (mean and covariance).

The two-dimensional formulation can be deemed effective if the following criteria are satisfied:

1. For the **joint optimal density**: its two major statistical central moments

(mean and covariance) should be close to the counterparts of the true density. It is also important that the estimated σ_{xy} in the covariance matrix should have the same sign (positive or negative) with the counterparts of the true density.

2. For the **joint maximum variance density**: while maintaining good estimations on the mean and the sign of σ_{xy}, the joint maximum variance density should have reasonably larger covariance (larger absolute value of all of σ_{xx}, σ_{xy} and σ_{yy}) compared with the joint optimal density.

Three 2-D distributions with different characteristics are selected as the true densities. The reasons for their selection are stated below:

- **Weak Linear Correlation Between x and y (negative)**: when the correlation between x and y is not strong (but still deemed as linearly correlated by Algorithm 2), the estimations from the 2-D formulation must be able to capture the weak correlation and give good estimations on the moments.

- **Strong Linear Correlation between x and y (Positive)**: when the correlation between x and y is strong, the strong linear correlation can surely be captured by the copula. Still, we need to examine the properties of such estimations in order to give comprehensive evaluations.

- **Irregular Linear Correlation between x and y (Positive)**: even if the correlation between x and y is deemed linear by Algorithm 2, the true joint density may not follow a line shape, making the estimation more challenging. We need at least one test case to see the performance of the 2-D formulation on such irregular joint densities.

3.7.1 Test Case 1: Weak Linear Correlation Between x and y (negative)

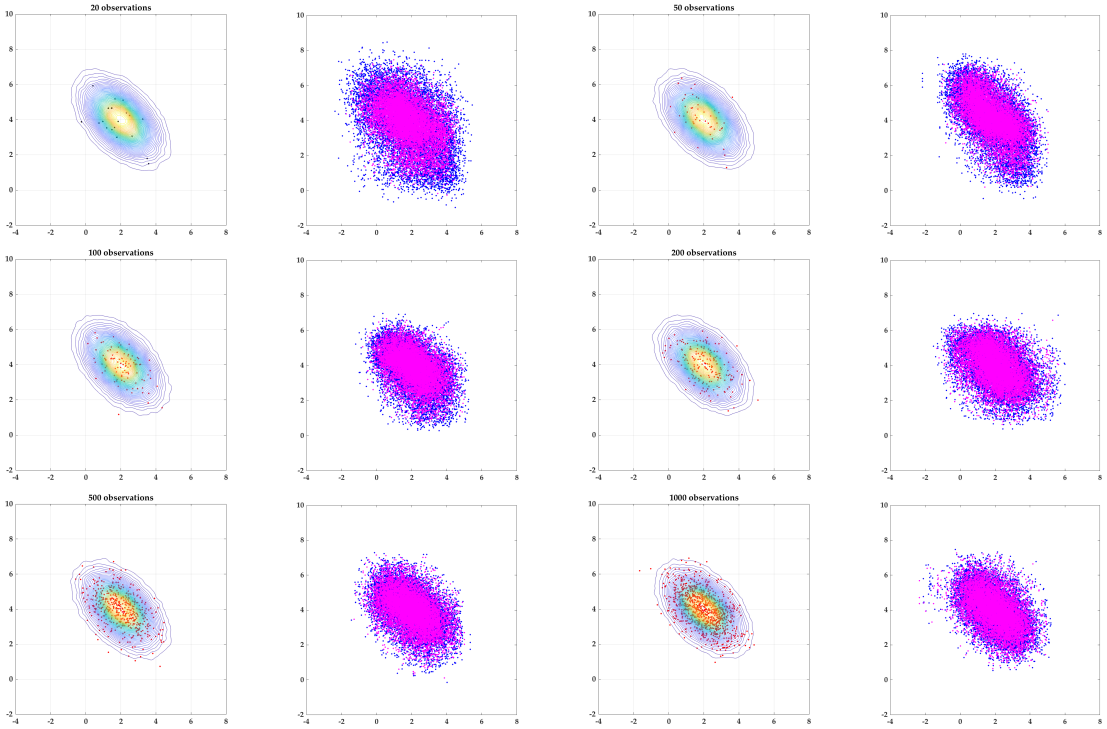


Figure 3.15: Test Case 1: weak linear correlation between x and y (negative) from 20 to 1,000 observations

Table 3.2: Test Case 1: comparison of the moments among the three densities

20 Data	**Ori.**	**Opt.**	**Max Var.**	**200 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean: μ_x	2	1.79	1.82	***Mean:*** μ_x	2	2.08	2.08
Mean: μ_y	4	3.89	3.75	***Mean:*** μ_y	4	3.95	3.92
Cov: σ_{xx}	1	1.18	2.12	***Cov:*** σ_{xx}	1	1.06	1.38
Cov: σ_{xy}	-0.5	-0.53	-1.05	***Cov:*** σ_{xy}	-0.5	-0.41	-0.53
Cov: σ_{yy}	1	1.33	2.72	***Cov:*** σ_{yy}	1	0.98	1.25
50 Data	**Ori.**	**Opt.**	**Max Var.**	**500 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean: μ_x	2	1.82	1.83	***Mean:*** μ_x	2	2.01	1.99
Mean: μ_y	4	4.19	4.09	***Mean:*** μ_y	4	3.99	3.99
Cov: σ_{xx}	1	1.03	1.45	***Cov:*** σ_{xx}	1	1.05	1.22
Cov: σ_{xy}	-0.5	-0.65	-1.02	***Cov:*** σ_{xy}	-0.5	-0.47	-0.55
Cov: σ_{yy}	1	1.24	2.06	***Cov:*** σ_{yy}	1	0.97	1.17
100 Data	**Ori.**	**Opt.**	**Max Var.**	**1000 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean: μ_x	2	2.11	2.11	***Mean:*** μ_x	2	1.97	1.97
Mean: μ_y	4	3.89	3.81	***Mean:*** μ_y	4	3.99	3.98
Cov: σ_{xx}	1	0.88	1.21	***Cov:*** σ_{xx}	1	1.02	1.18
Cov: σ_{xy}	-0.5	-0.39	-0.56	***Cov:*** σ_{xy}	-0.5	-0.52	-0.59
Cov: σ_{yy}	1	0.85	1.22	***Cov:*** σ_{yy}	1	1.07	1.23

3.7.2 Test Case 2: Strong Linear Correlation between x and y (Positive)

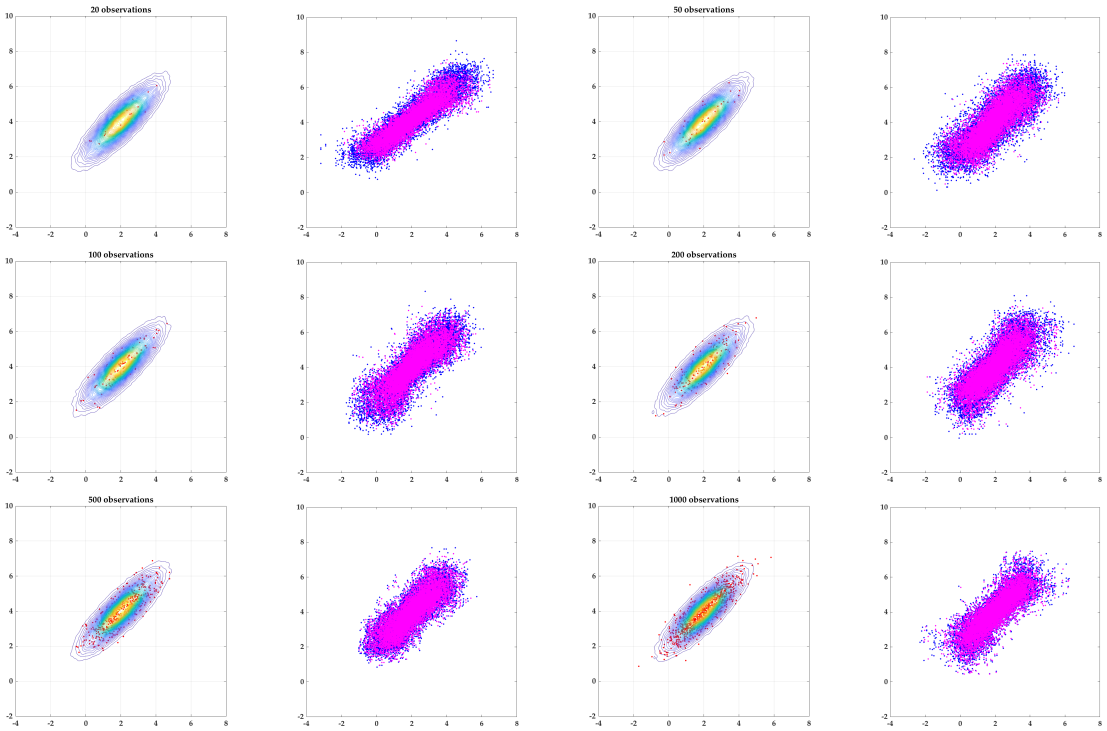


Figure 3.16: Test Case 2: strong linear correlation between x and y (Positive) from 20 to 1,000 observations

Table 3.3: Test Case 2: comparison of the moments among the three densities

20 Data	**Ori.**	**Opt.**	**Max Var.**	**200 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean: μ_x	2	2.01	2.03	***Mean:*** μ_x	2	1.99	1.99
Mean: μ_y	4	4.18	4.25	***Mean:*** μ_y	4	4.02	4.06
Cov: σ_{xx}	1	1.73	2.81	***Cov:*** σ_{xx}	1	1.11	1.41
Cov: σ_{xy}	0.8	1.19	2.02	***Cov:*** σ_{xy}	0.8	0.92	1.17
Cov: σ_{yy}	1	1.01	1.74	***Cov:*** σ_{yy}	1	1.22	1.57
50 Data	**Ori.**	**Opt.**	**Max Var.**	**500 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean: μ_x	2	1.91	1.93	***Mean:*** μ_x	2	2.03	2.04
Mean: μ_y	4	3.99	4.01	***Mean:*** μ_y	4	3.99	4.01
Cov: σ_{xx}	1	1.22	1.92	***Cov:*** σ_{xx}	1	1.08	1.23
Cov: σ_{xy}	0.8	0.97	1.48	***Cov:*** σ_{xy}	0.8	0.87	0.99
Cov: σ_{yy}	1	1.21	1.75	***Cov:*** σ_{yy}	1	1.06	1.21
100 Data	**Ori.**	**Opt.**	**Max Var.**	**1000 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean: μ_x	2	2.04	2.07	***Mean:*** μ_x	2	2.02	2.03
Mean: μ_y	4	4.04	4.01	***Mean:*** μ_y	4	4.01	4.03
Cov: σ_{xx}	1	1.19	1.65	***Cov:*** σ_{xx}	1	1.12	1.23
Cov: σ_{xy}	0.8	0.99	1.41	***Cov:*** σ_{xy}	0.8	0.92	1.01
Cov: σ_{yy}	1	1.21	1.72	***Cov:*** σ_{yy}	1	1.11	1.22

3.7.3 Test Case 3: Irregular Linear Correlation between x and y (Positive)

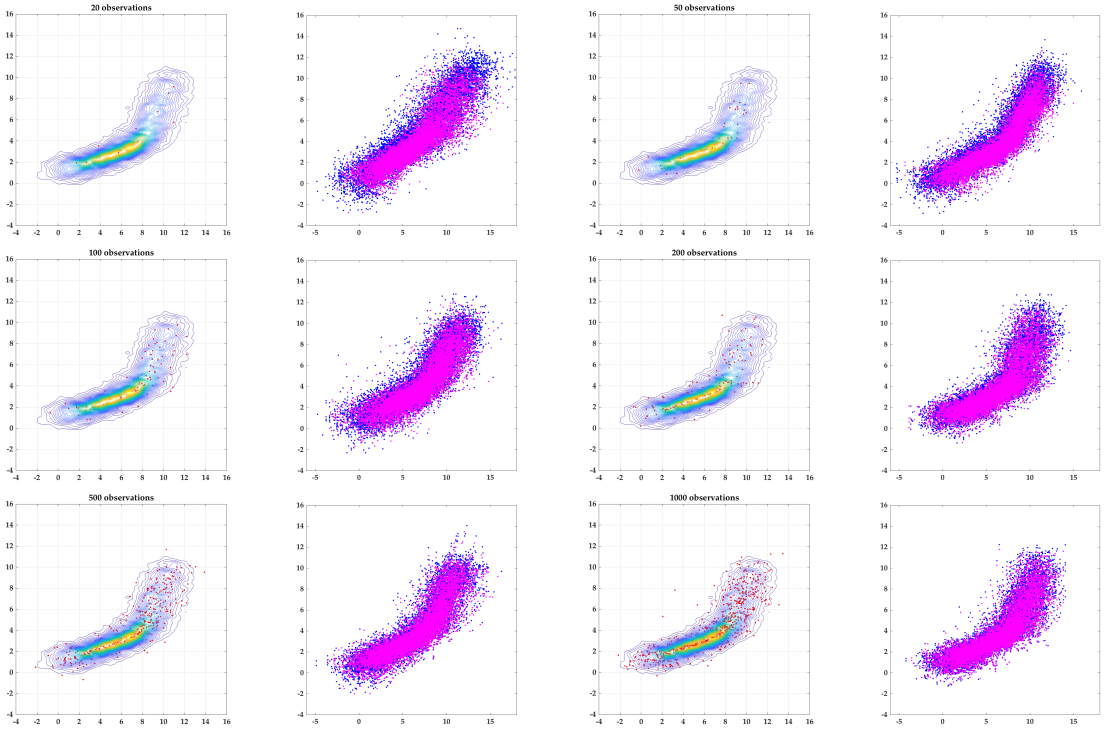


Figure 3.17: Test Case 3: irregular linear correlation between x and y (Positive) from 20 to 1,000 observations

Table 3.4: Test Case 3: comparison of the moments among the three densities

20 Data	**Ori.**	**Opt.**	**Max Var.**	**200 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean: μ_x	6.37	6.25	6.47	***Mean:*** μ_x	6.37	6.11	6.04
Mean: μ_y	4.05	3.91	4.77	***Mean:*** μ_y	4.05	3.85	4.15
Cov: σ_{xx}	9.01	9.84	16.17	***Cov:*** σ_{xx}	9.01	8.75	10.08
Cov: σ_{xy}	5.81	6.49	11.45	***Cov:*** σ_{xy}	5.81	5.27	6.96
Cov: σ_{yy}	5.49	5.68	10.17	***Cov:*** σ_{yy}	5.49	4.94	6.58
50 Data	**Ori.**	**Opt.**	**Max Var.**	**500 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean: μ_x	6.37	6.43	5.96	***Mean:*** μ_x	6.37	6.41	6.25
Mean: μ_y	4.05	4.06	4.44	***Mean:*** μ_y	4.05	4.05	4.17
Cov: σ_{xx}	9.01	10.71	15.01	***Cov:*** σ_{xx}	9.01	10.18	11.68
Cov: σ_{xy}	5.81	7.15	10.46	***Cov:*** σ_{xy}	5.81	6.68	6.68
Cov: σ_{yy}	5.49	6.17	9.08	***Cov:*** σ_{yy}	5.49	5.98	5.98
100 Data	**Ori.**	**Opt.**	**Max Var.**	**1000 Data**	**Ori.**	**Opt.**	**Max Var.**
Mean: μ_x	6.37	7.01	6.72	***Mean:*** μ_x	6.37	6.25	6.21
Mean: μ_y	4.05	4.33	4.61	***Mean:*** μ_y	4.05	3.89	4.05
Cov: σ_{xx}	9.01	9.25	12.47	***Cov:*** σ_{xx}	9.01	9.78	10.56
Cov: σ_{xy}	5.81	5.94	8.16	***Cov:*** σ_{xy}	5.81	6.02	6.83
Cov: σ_{yy}	5.49	5.45	7.32	***Cov:*** σ_{yy}	5.49	5.32	6.29

3.8 Results from the Test Cases

Results from the three test cases show that the 2-D formulation can effectively model dependency between x and y and estimate the true density in different situations. Below we summarize the observations made from the test cases:

1. **Test Case 1 & 2**: for both weak and strong linear correlation cases, the joint optimal densities can provide good estimations for the true densities, as their means and covariances are close to the true values. The joint maximum variance densities are effective in providing reasonably larger covariances (in all of σ_{xx}, σ_{xy} and σ_{yy}) while still maintaining good estimations on the means. Even in the weak linear correlation case, both the joint optimal density and the joint maximum variance density can catch the correct correlation type (positive or negative). As the dataset size grows, the joint maximum variance densities approach to the joint optimal densities as expected.

2. **Test Case 3**: the irregular linear correlation case here is a joint distribution with a crescent shape. Although having a strong linear correlation between x and y, the curve of this true density still creates difficulties (in the form of covariance) for the copula-based formulation to estimate. In such a situation, the results are still promising. As can be seen from Table 3.4, even with less than 100 data, the joint optimal densities and the joint maximum variance densities can successfully estimate the properties of the true covariance (such as the rank in magnitude among σ_{xx}, σ_{xy} and σ_{yy}). The results show that for such cases, the 2-D formulation can still effectively estimate the true density, and is very promising in the uncertainty propagation process.

More in-depth comparisons, and descriptions on observed limitations of the 2-D formulation are in the following two sections.

3.9 Comparison with the t-copula

It is worth comparing the two-dimensional solution (NP + Frank copula) with the t-copula, because the t-copula can be thought of as representing the dependence structure implicit in a multivariate t-distribution [52], which also has the uncertainty consideration in it. Compared with the unique copula of a multivariate Gaussian distribution, it is a limiting case of t-copula as $\nu \to \infty$.

The t-copula has received much recent attention, and is thought to be superior to the Gaussian copula, because of its ability to capture better the phenomenon of dependent extreme values. In this section we compare some effects of the NP + Frank copula and the t-copula in two situations: weak linear correlation and very strong linear correlation.

3.9.1 Performances in Weak Linear Correlation

In this test, a multivariate Gaussian distribution with a PCC of 0.4 is used as the true density, where a small dataset is sampled. Then both formulations are applied to the small dataset to capture the dependence and estimate the true density.

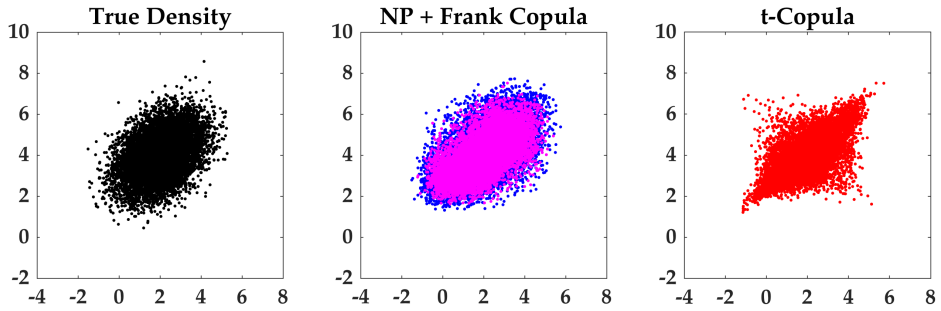


Figure 3.18: Comparison between the NP + Frank copula and the t-copula on the same small dataset from weak linear correlation ($PCC = 0.4$)

Figure 3.18 shows that in cases like this when x and y are in a weak linear correlation, estimation from the t-copula may have a star-like shape like the right graph in

the Figure. Compared with the true density and the NP + Frank copula estimation, the overall results demonstrate that in such weak linear correlation cases, the NP + Frank copula estimation is more stable and a better choice than the t-copula.

3.9.2 Performances in Very Strong Linear Correlation

A similar test is conducted using a multivariate Gaussian distribution with a PCC of 0.95 as the true density. Figure 3.19 gives the results from the two estimations. It can be found that under the strong linear correlation case, both estimations can well capture the dependence between x and y. But when looking into some details, two tails of the NP + Frank copula estimation tend to be a little bit divergent which can not properly reflect the shape of the true density near the extreme values. This is because the points near extreme values are always more sparse, and therefore are deemed more 'uncertain' by the two-dimensional formulation in this thesis. Under such very strong linear correlation cases, the t-copula performs better.

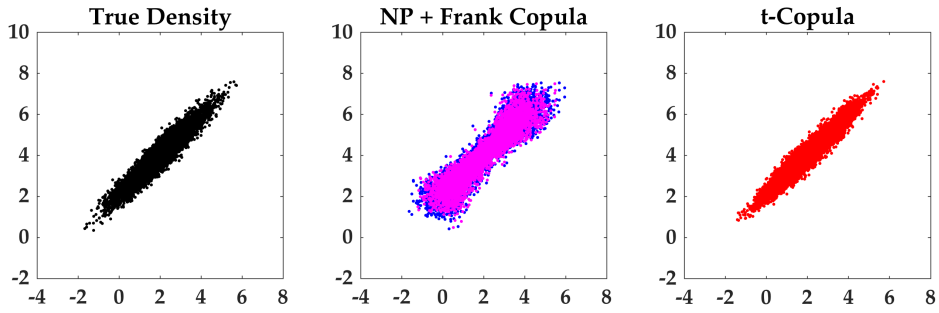


Figure 3.19: Comparison between the NP + Frank copula and the t-copula on the same small dataset from very strong linear correlation ($PCC = 0.95$)

As a result, NP + Frank copula is recommended in the weak linear correlation case, and t-copula is recommended in the very strong linear correlation case.

3.10 Performance on 2-D Mixture Model

Another interesting study for the two-dimensional estimations is its effectiveness on 2-D mixture models. Since copulas is the core in the 2-D estimations and its main function is to capture the dependency between the variables, it is questionable whether such a copula-based formulation can estimate multimodal properties displayed by mixture models. Therefore, in this study, we apply the 2-D estimations on samples from a mixture model consists of two multivariate Gaussian distributions, as shown in Figure 3.20.

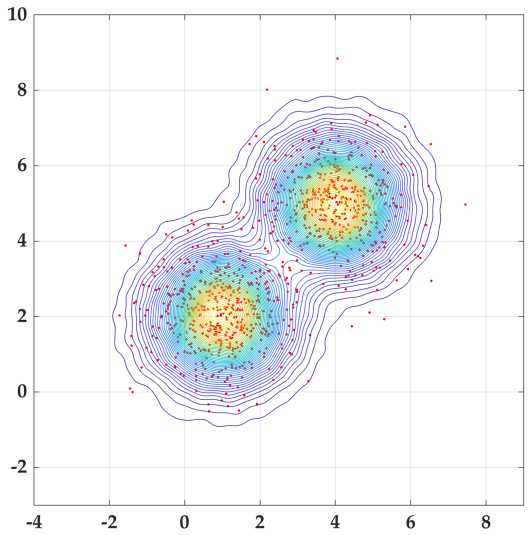


Figure 3.20: A 2-D mixture model consists of two multivariate Gaussian distributions

Results of the study is shown in Figure 3.21. Experimental results show that the copula-based 2-D estimations do have capability of estimating mixture models that have more than one mode (peak) when we have enough data (say, $n > 200$). However, even with more than 1000 observations here, it may not recover the original mixture model as good as the multivariate kernel density estimations. An important reason for this is that the copula-based methods are usually **not estimating the mixture model from the principal component direction**. For example, the principal component direction for the mixture model in Figure 3.20 is $[1, 1]$, on which

the data plotted can best display the difference between the two peaks. Nevertheless, the copula-based nonparametric formulation first estimates marginal densities in the x ($[1, 0]$) and y ($[0, 1]$) directions separately. It can be imagined that when projected on either $[1, 0]$ or $[0, 1]$, the multimodal property of the mixture model is less distinguishable. The small dataset case will make estimating the multimodal property even more difficult.

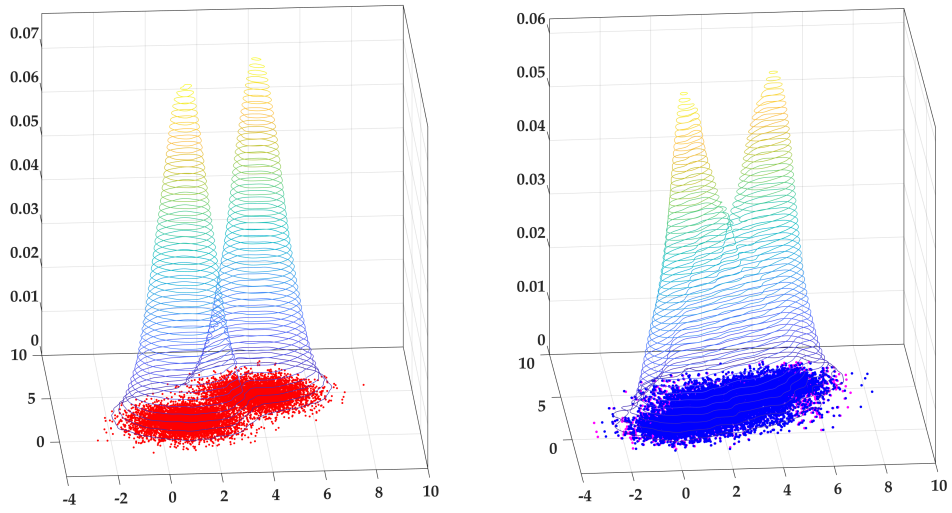


Figure 3.21: Comparison of the an original mixture model consists of two multivariate Gaussian distributions (left) and its 2-D estimations using copula (right)

With these observations, the author recommend the use of multivariate kernel density estimation when there is multimodal property in the 2-D data. But then again, in science and engineering applications, even if correlation exist among variables, it is very rare that their joint densities have multimodal properties like this. From this angle, the copula-based 2-D estimation can still be used (and enough) to solve most of the real-world problems.

CHAPTER 4
EXPERIMENTS ON UNCERTAINTY PROPAGATION AND THE AEDT

4.1 A Complete Uncertainty Propagation Test

4.1.1 Problem Set-up

The complete uncertainty propagation test is designed to examine the effectiveness of both the one-dimensional and the two-dimensional nonparametric formulations on propagating uncertainty due to small datasets. A representative surrogate model in systems engineering is first used in this uncertainty propagation exercise.

In systems engineering, surrogate models are often used to save computational costs. As introduced in the previous chapters, the response surface equations (RSE) and the artificial neural network (ANN) are the two most commonly used surrogate models. For this test, a RSE that includes **9 input variables** (5 independent and 4 dependent) is created, all of which are uncertainty sources with some certain distributions. This RSE contains three main components: **first-degree polynomial** ($a_i \cdot x_i, i = 1, ..., 9$), **second-degree polynomial** ($b_i \cdot x_i^2, i = 1, ..., 9$), and **second-degree factorial** ($c_i \cdot x_i x_j, i, j = 1, ..., 9$). Details on the true distributions for each input variable can be found in Table 4.1.

Samples with six different sizes are used in this test, using the following procedure:

1. **Step 1 - Create Small Datasets**: for each uncertainty propagation experiment, create random samples of size n from the true distributions of the input variables as the small datasets

2. **Step 2 - Generate Nonparametric Density Estimations**: for each in-

Table 4.1: True distributions of the 9 input variables in the complete uncertainty propagation test

Input Name	Dependency Status	True Distribution
x_1	Independent	$X \sim Lognormal(5.0, 0.30^2)$
x_2	Independent	$X \sim Lognormal(4.0, 0.20^2)$
x_3	Independent	$X \sim Normal(300, 40^2)$
x_4	Independent	$X \sim Gamma(80, 2.5)$
x_5	Independent	$X \sim Weibull(180, 6)$
x_6	Dependent	x of $X \sim MVN([100, 120], [1\ 0.85; 0.85\ 1])$
x_7	Dependent	y of $X \sim MVN([100, 120], [1\ 0.85; 0.85\ 1])$
x_8	Dependent	x of $X \sim MVN([60, 80], [1\ -0.5; -0.5\ 1])$
x_9	Dependent	y of $X \sim MVN([60, 80], [1\ -0.5; -0.5\ 1])$

dependent input variable, apply the 1-D formulation to estimate the optimal sampling density and the maximum variance density. For the correlated input variables, apply the 2-D formulation to estimate the joint optimal density and the joint maximum variance density, based on small datasets from Step 1.

3. **Step 3 - Uncertainty Propagation through Systems Model**: using Monte Carlo Simulation (MCS), generate random samples of 10,000 observations from the estimated densities mentioned in Step 2, and obtain the output distributions. The optimal output distribution is generated by propagating the optimal sampling densities, and the maximum variance output distribution is generated by propagating the maximum variance sampling densities.

4. **Step 4 - Repeat the Process**: change the n from 20 to 50, 100, 200, 500, and 1,000, and repeat Steps 1-3 for six times to finally obtain six groups of uncertainty propagation results.

5. **Step 5 - Compare the Results**: uncertainty propagation result from each group (with certain sample size n) is then compared with the result obatined by propagating the true densities. Then observe the patterns to examine the effectiveness of the 1-D and 2-D formulations.

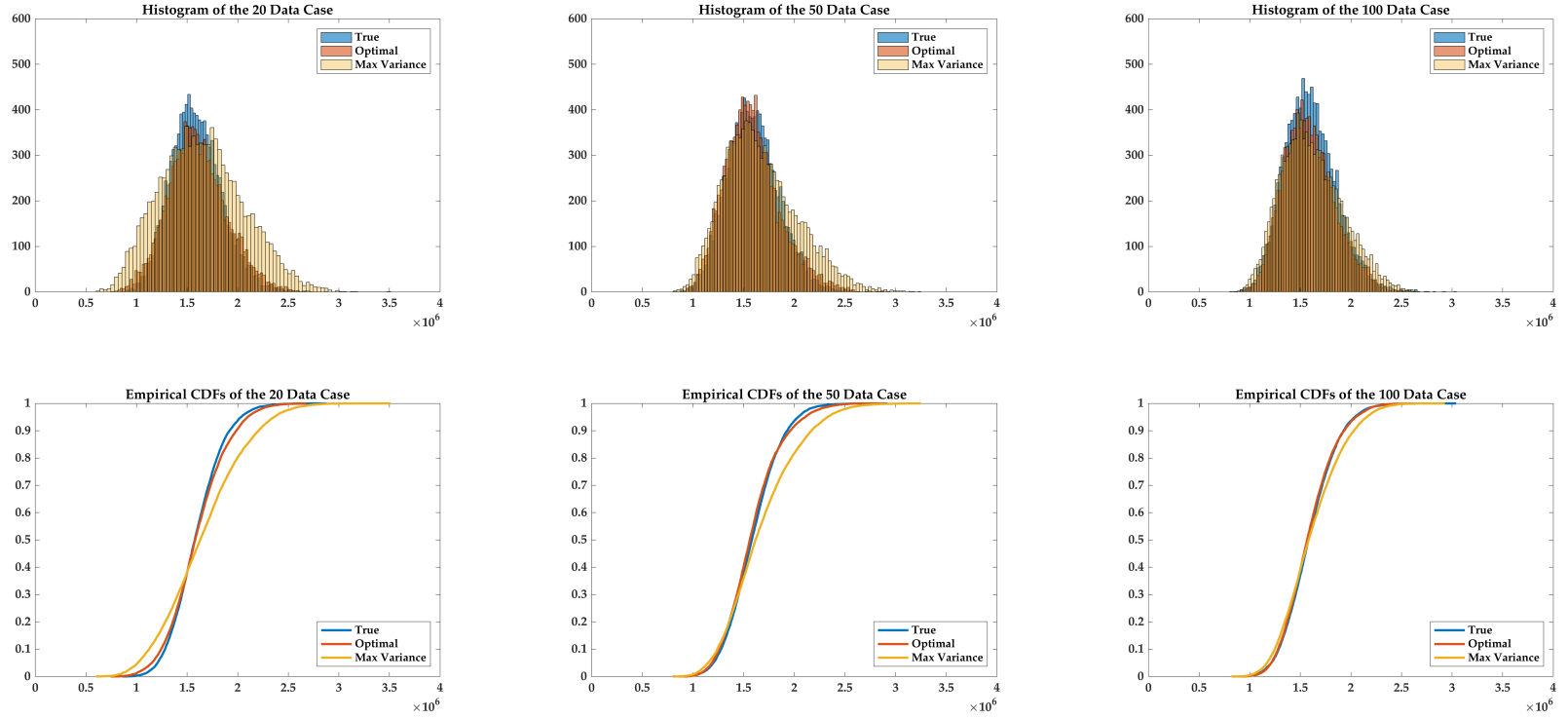


Figure 4.1: Uncertainty propagation results Part I: output histograms and empirical CDFs for 20 observations (left), 50 observations (middle), and 100 observations (right)

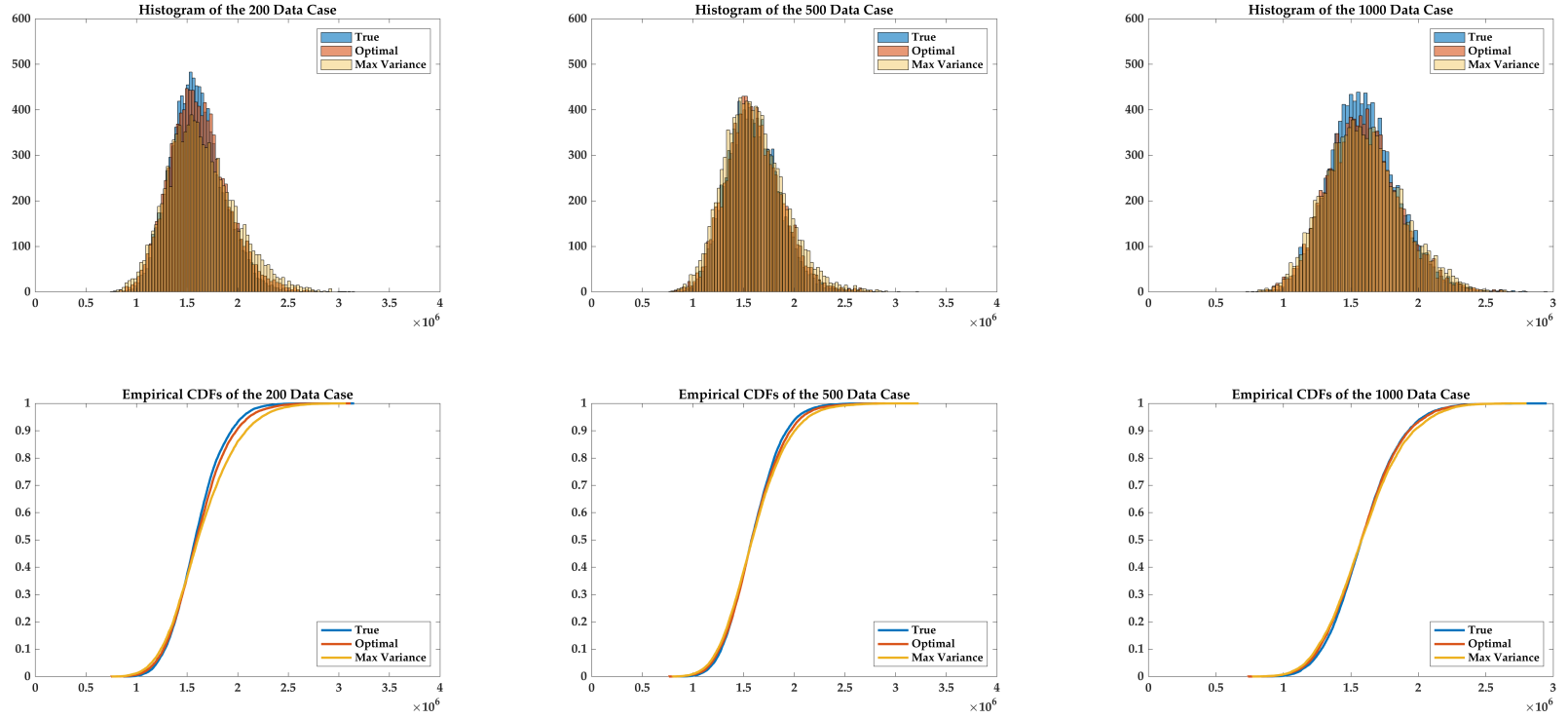


Figure 4.2: Uncertainty propagation results Part II: output histograms and empirical CDFs for 200 observations (left), 500 observations (middle), and 1000 observations (right)

4.1.2 Uncertainty Propagation Results

Figure 4.1 and 4.2 contain the results from the uncertainty propagation test. The 1-D and 2-D formulations can be deemed effective through the following observations:

1. Even when the sample sizes $n < 100$, the optimal estimations for the output (both histograms and empirical CDFs, colored in orange) are very close to the those of the true output (colored in blue). This shows that when we propagate the optimal sampling densities (1-D and 2-D) through MCS, distribution of the output can be well estimated.

2. For the maximum variance distributions for the output, the histograms show that they are indeed more 'widespread', and can represent results with more uncertainties than the optimal results. The difference can also be observed through the empirical CDFs, and this is what we expected from propagating the maximum variance densities (1-D and 2-D). The maximum variance results provides a way to capture more potential extreme values in the output, and can help us to make safer decisions under the influence of small datasets.

3. As the sample size grows from 20 to 1,000, the maximum variance results (histograms and empirical CDFs) are getting closer to the optimal results. In the mean time, both of them are getting closer to the true result. This shows that when we gradually have more data to estimate the true distribution of the uncertainty inputs, less uncertainty is included in the density estimation process. As a result, the nonparametric 1-D and 2-D formulations can better recover the distribution of the inputs, leading to better estimations of the output.

4.1.3 Comparison with the Independent Assumption

Since the 2-D copula-based formulation is motivated by the fact that output uncertainty can be **overestimated or underestimated** when correlations among input

variables are ignored, a comparison with results from independent assumption is conducted. In order to enlarge the differences between the two results and illustrate the effect of dependency modeling, the systems model was modified to emphasize the role of correlated input variables.

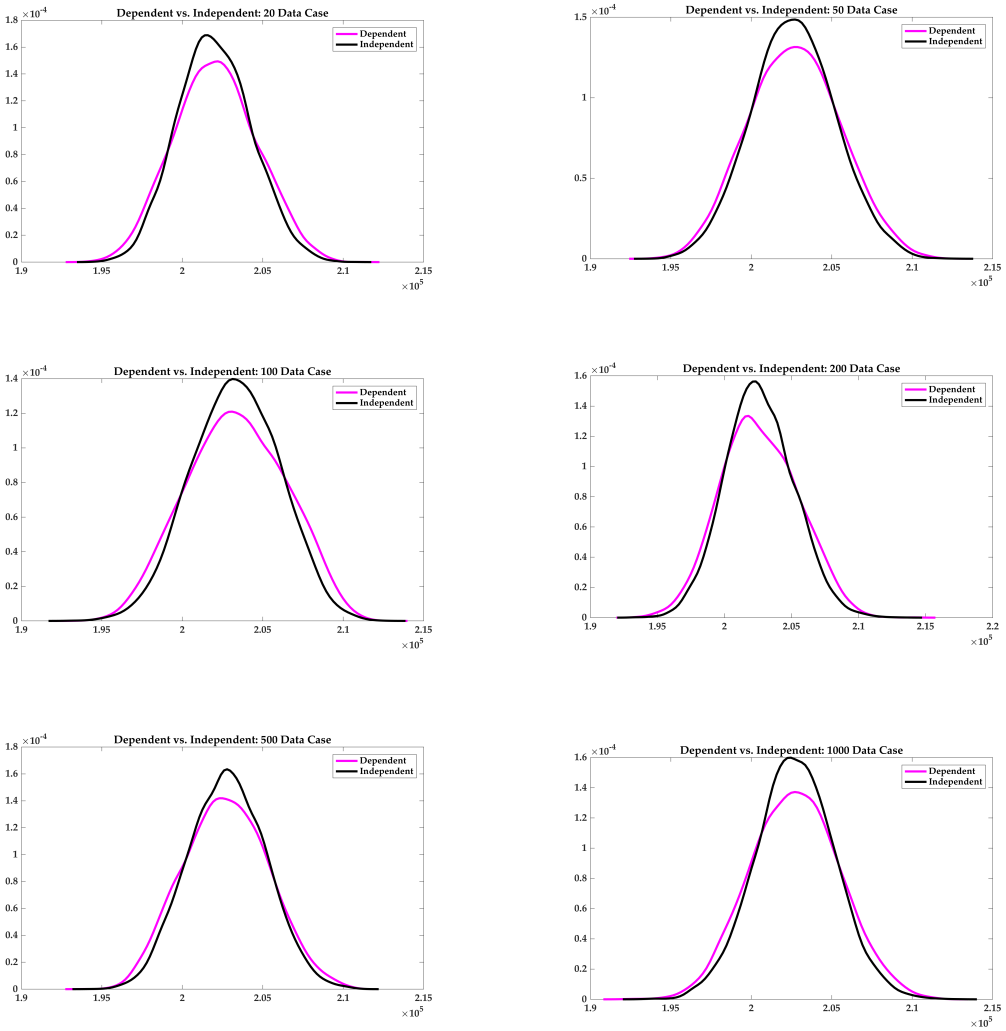


Figure 4.3: Comparison between independent and dependent assumptions: from 20 to 1,000 observations

Figure 4.3 shows the results of the comparison. At each sample size, the output distribution under the independent assumption has smaller variance compared to the output distribution obtained through dependence modeling. This means that for this systems model, uncertainty in the output can be underestimated when correlations

among some inputs are ignored. This again demonstrates the necessity of using multivariate formulation to capture dependencies when propagating uncertainty through a systems model.

4.2 Test on an Aviation Environmental Impact Analysis

4.2.1 Airline Data

The Aviation Environmental Design Tool (AEDT) is used to assess fleet-wide fuel burn, emissions, and noise impacts. There are three main outputs for the AEDT: **Fuel Burn**, **Emission**, and **Noise**. Even though there are many inputs for the tool, this study only investigate the two most important ones: **Engine Thrust** and **Takeoff Gross Weight**, in the form of percentage (%) of their maximum. In general, distributions of % engine thrust (% Thrust) and % takeoff gross weight (% GW) are relatively complicated. Figure 4.4 and 4.5 show the data for Boeing 737-800 aircraft for more than 60,000 flights from an airline.

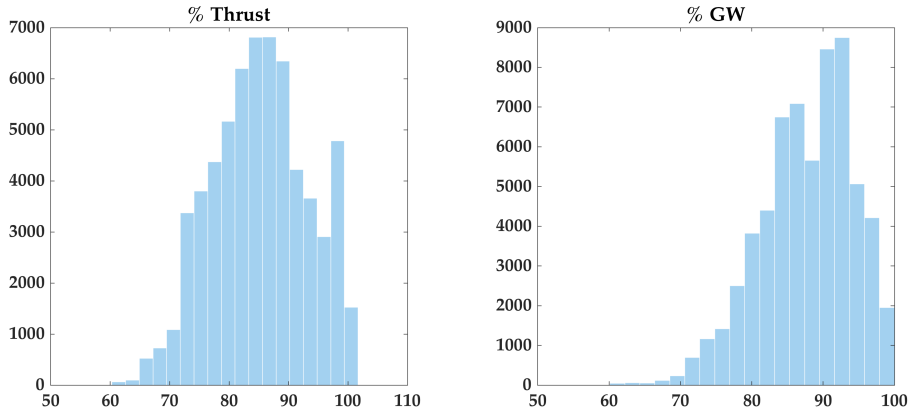


Figure 4.4: Histograms of the marginal distributions of % Thrust and % GW for Boeing 737-800 aircraft (over 60,000 flights)

We need to pay special attention to the joint distribution of % Thrust and % GW for this group of data shown in Figure 4.5. The Figure shows that this joint

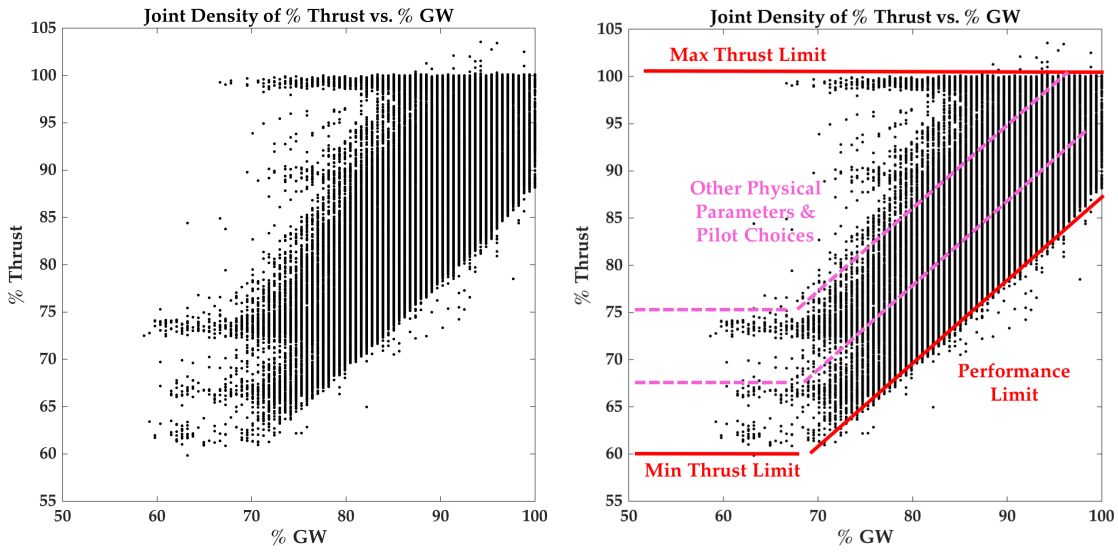


Figure 4.5: Scatter plot of % Thrust and % GW for Boeing 737-800 aircraft (over 60,000 flights)

density is sophisticated and impossible to be described by any common joint densities. Due to the rules in aircraft performance, it is safe to say that % Thrust and % GW are correlated (positive correlation) as thrust needed generally increase with the takeoff weight. However, apart from that, patterns of this joint distribution are also influenced by other factors and bounds:

- **Maximum Thrust Limit**: % Thrust can not exceed 100% normally
- **Minimum Thrust Limit**: % Thrust can not be lower than 60% due to a few performance considerations
- **Performance Limit**: influenced by factors such as the takeoff field length and the second segment climb
- **Other Physical Parameters and Pilot Choices**

4.2.2 Insufficient Data Case

Accurate assessments of fuel burn, emission and noise often require a relatively complete dataset that include a large amount of data integrated from different databases.

For example, only the Boeing 737-800 data from a single airline within a year has as many as 60,000 flights. Nevertheless, for most of the researchers who are in this field, it is impossible for them collect all these data in order to study a topic. Their are three common constraints in the process of data collection:

1. **Time**: sometime we can only acquire a dataset nationwide **for a specific day**: for example, on December 06, 2014

2. **Space**: sometime we can only acquire a dataset from **a specific airport**: for example, data from only ATL

3. **Time & Space**: we may only acquire a dataset from **a specific airport on a specific day**: for example, at JFK on June 21, 2014 (the most strict one)

In the meantime, flight data from each airport or single day are heterogeneous: each day, there are different number of flights from/to different places; at each airport, the number of flights and destinations varies. We can not speculate the whole dataset by simply scaling the small dataset up (even if we do that, how can we deal with the uncertainty within it?). In this case, we only know that the small dataset we have is a subset of the whole dataset, and we need to do our best to still do the assessment.

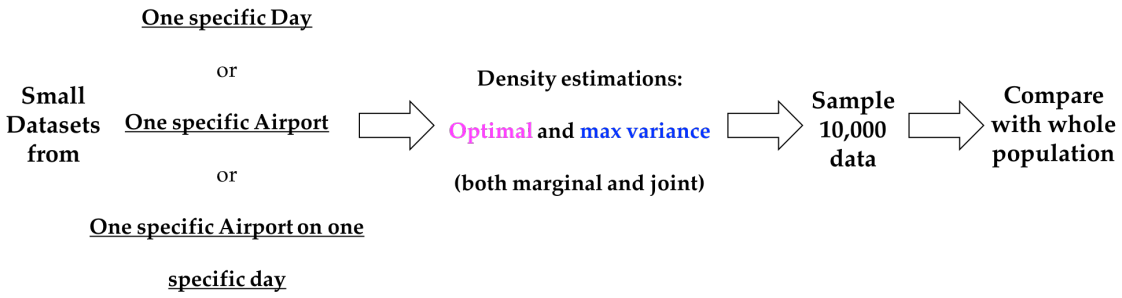


Figure 4.6: Process of estimating Boeing 737-800 data from small datasets

Even though the Boeing 737-800 data shown in Figure 4.4 and 4.5 is a complicated joint distribution with physical bounds and chaotic factors, we still apply the 1-D and 2-D formulations on it to make a few observations (test process in Figure 4.6).

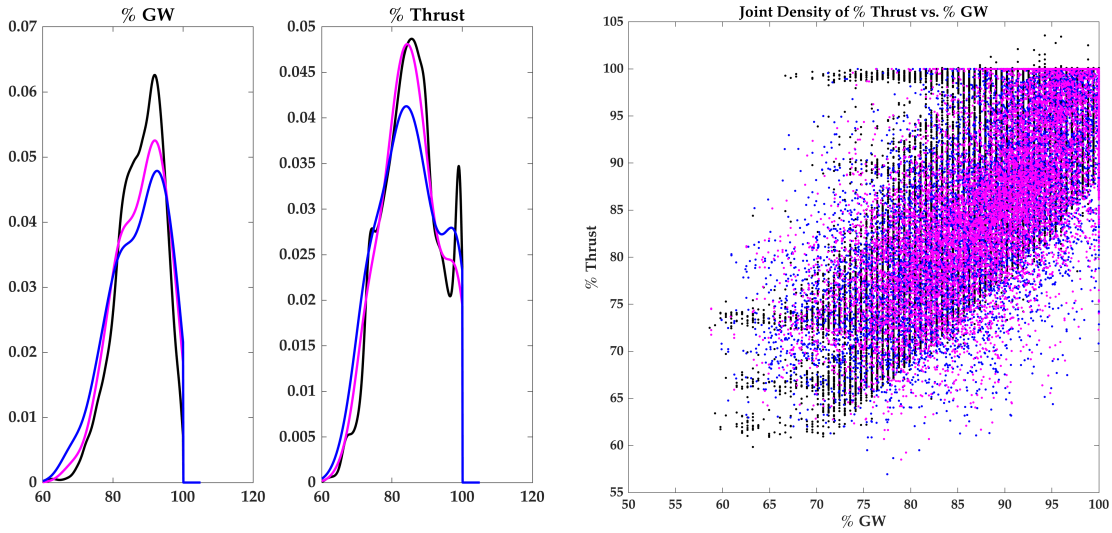


Figure 4.7: Optimal (magenta) and Maximum Variance (blue) Estimations with true density (black), estimated from Boeing 737-800 data on December 06, 2014, with 209 out of 62,493 data (0.33%)

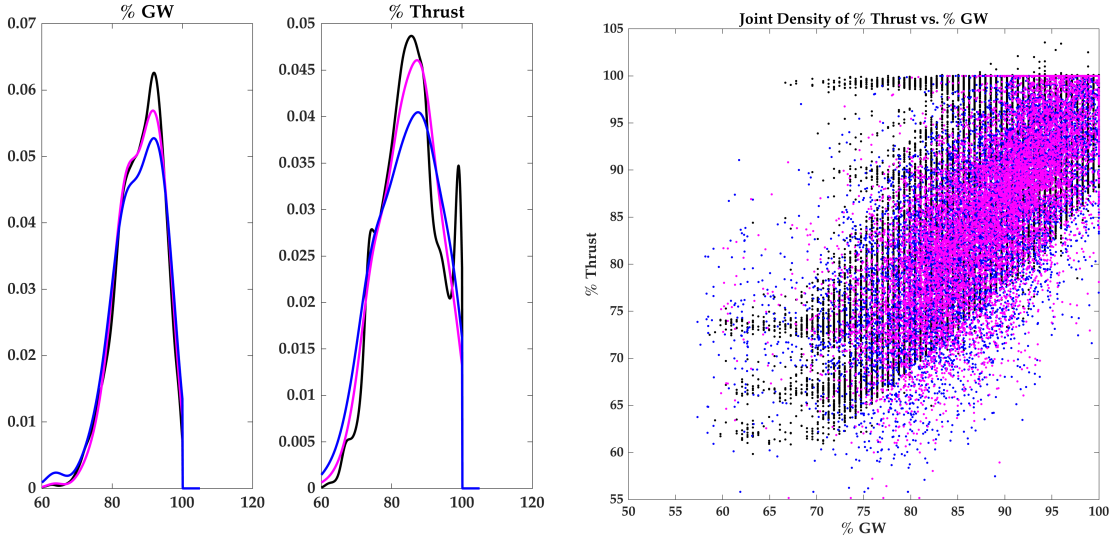


Figure 4.8: Optimal (magenta) and Maximum Variance (blue) Estimations with true density (black), estimated from Boeing 737-800 data on September 15, 2014, with 232 out of 62,493 data (0.37%)

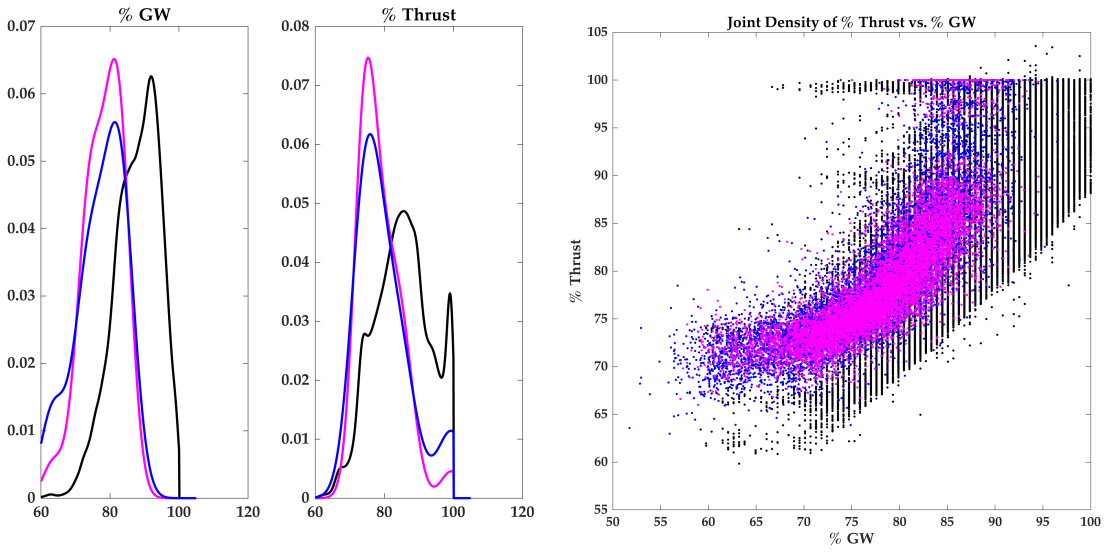


Figure 4.9: Optimal (magenta) and Maximum Variance (blue) Estimations with true density (black), estimated from Boeing 737-800 data at STL, with 59 out of 62,493 data (0.09%)

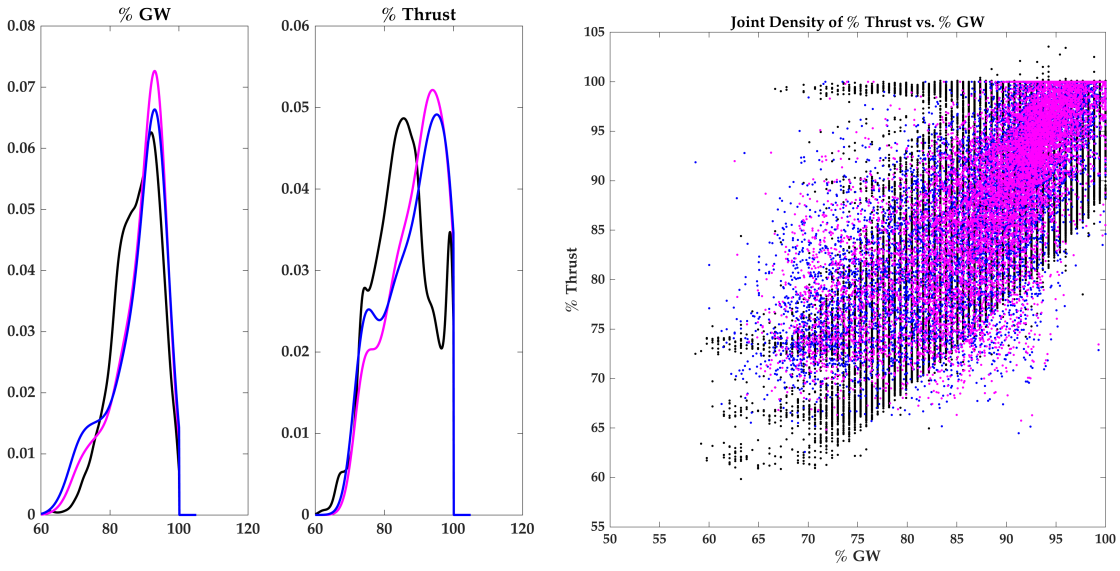


Figure 4.10: Optimal (magenta) and Maximum Variance (blue) Estimations with true density (black), estimated from Boeing 737-800 data at SFO, with 342 out of 62,493 data (0.55%)

Business Case 3: Time & Space Constraint

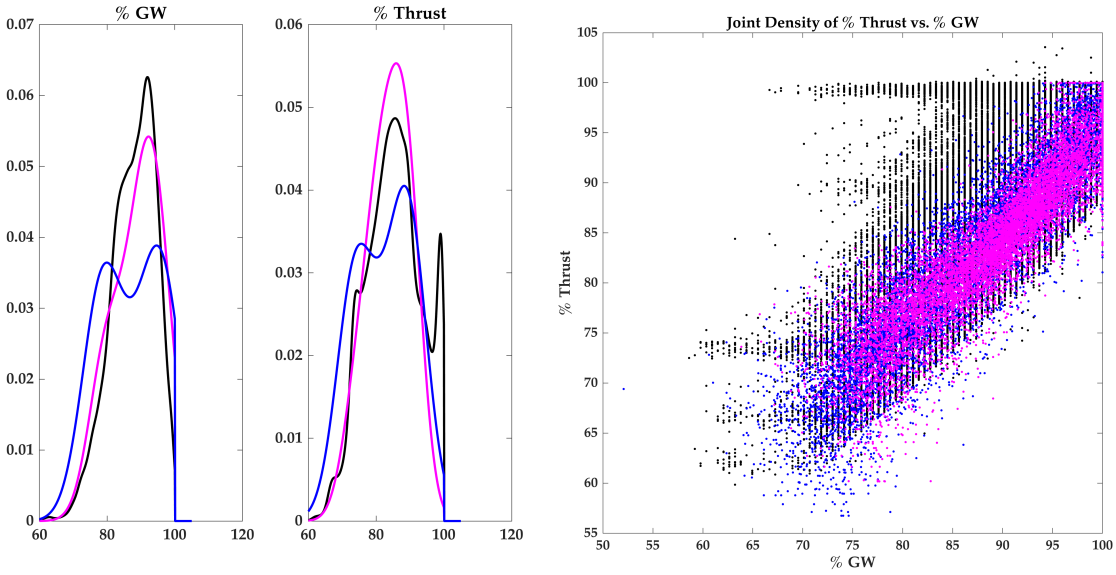


Figure 4.11: Optimal (magenta) and Maximum Variance (blue) Estimations with true density (black), estimated from Boeing 737-800 data at DTW on September 15, 2014, with 10 out of 62,493 data (0.02%)

Observations

- **Time Constraint Case**: Figure 4.7 and 4.8 together show that when using Boeing 737-800 data from a specific day, the whole population (both marginal and joint) can be well estimated.

- **Space Constraint Case**: Figure 4.9 and 4.10 together show that Boeing 737-800 data from a specific airport can not be used to effectively estimate the whole population. From the STL data (Figure 4.7), the estimated marginal densities for both % GW and % Thrust have relatively low mean values compared to the true marginal distributions. The estimated joint densities can also partially cover the the area of the true joint density. These show that data from STL is biased compared with the whole population. Similar observations can also

be made for SFO data (Figure 4.8), in which the estimated marginal densities for both % GW and % Thrust have higher mean values compared to the true marginal distributions. In the end, estimations from SFO data fail to recover the complete picture of the whole population.

- **Time & Space Constraint Case**: in fact, after we conclude that data from a specific airport can not be used to properly estimate the whole population, results from this stricter case can be imagined. Figure 4.11 shows that when using data at DTW airport on a specific day, estimated joint densities can also only cover part of the true joint distribution, making the estimations ineffective.

With those observations, a recommendation is that when we actually only have limited time or resources to collect data, then in cases like this, it would be better to collect data based on a time unit (one specific day), not space unit (one airport). Data collected based on time unit reflects operations across geographical limits and is therefore more representative.

To explain why data from a specific airport may fail to reflect operations across the nation, a brief study was conducted. Among the 62,493 flights, counts for the departure airports are in Table 4.2.

Table 4.2: Airport counts (departure) from the 62,493 Boeing 737-800 flights

Airport	**Count**	**Airport**	**Count**
JFK	9,612	LAS	2,446
ATL	7,356	BOS	1,780
LAX	5,016	...	...
SLC	4,783	SFO	342
DTW	3,004	...	...
MSP	2,764	ORD	67
SEA	2,702	MDW	17

The airport statistics shows that when this is just Boeing 737-800 flights from a single airline, then data from a specific airport may not be persuasive sample. For

example, two airports from Chicago (ORD and MDW) actually have a lot of 737-800 flights, yet this dataset only records as few as 67 and 17 flights respectively. Also, each airline has its own operations characteristics: in the headquarter airports, they have dense flights; in some other airports, they may have less flights but the flights are always full (resulting heavier takeoff weights).

4.2.3 AEDT Uncertainty Propagation Test

In the uncertainty propagation test on AEDT, we study fuel burn and emissions for Boeing 737-800 at Atlanta airport (ATL). The flights that depart from ATL are extracted, resulting a population of 7,344 flights, plotted in Figure 4.12. From this new population, we randomly select data from four specific days, which forms four small sample with sample sizes ranging from 24 to 38. For each small sample, the 2-D nonparametric formulation developed in this thesis is applied to estimated the joint density of the population at ATL. New samples generated from the estimated densities are used as the inputs for AEDT, and the corresponding output distributions are compared to the true distributions obtained from the complete ATL dataset.

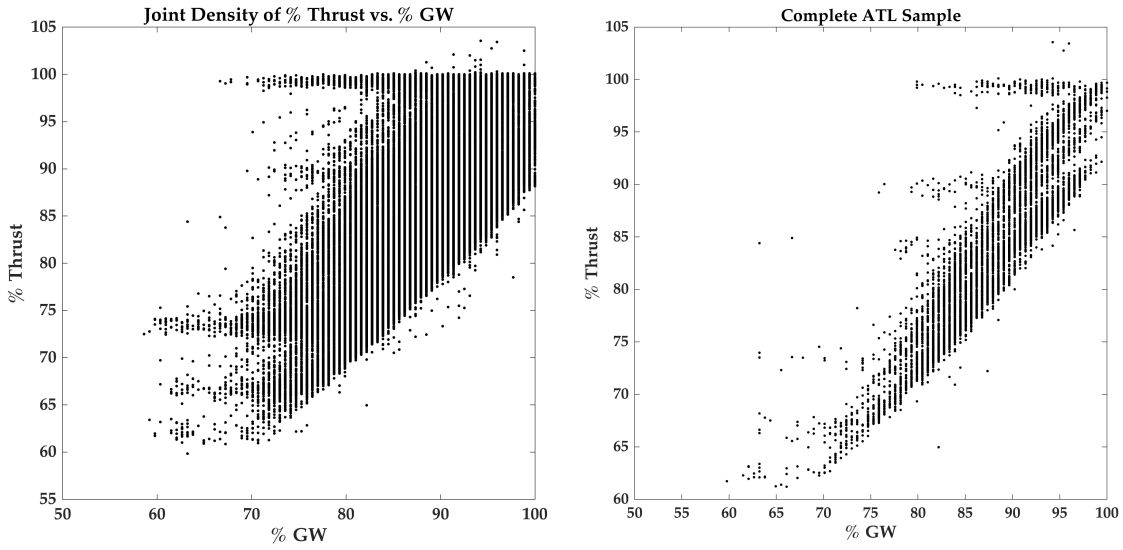


Figure 4.12: Comparison of scatter plots of % Thrust and % GW for Boeing 737-800 aircraft: complete version (left) and ATL only (right)

4.2.4 AEDT Test Results

In this uncertainty propagation test through AEDT, a study is conducted to examine: *"How well can we estimate the distributions of fuel burn and emissions at ATL throughout a year by only using data from one day?"*. In fact, the AEDT has three main outputs: fuel burn, emission, and noise (contour around airport). Considering the extensive amount of time needed to run the noise analysis (60 times compared with justing analyzing fuel burn and emission), only the fuel burn and the emission are analyzed here.

To start with, flights' data from four specific days are randomly selected from the whole dataset, which ranges from March 2014 to March 2015. The four days are: September 25, 2014, June 22, 2014, December 15, 2014, and February 27, 2015. Compared to the complete dataset which has 7,344 flights, dataset of these four days have 37 flights, 38 flights, 26 flights and 24 flights respectively. Before the uncertainty propagation through AEDT, both the 1-D and 2-D nonparametric estimations are applied on the four small datasets which contains GW% and Thrust% data, and the marginal and joint density estimations of the four small datasets are shown in Figure 4.13. It can be observed that compared with the true marginal and joint densities (in black), both the optimal density and maximum variance density in each case can provide valid estimations of the true density. Besides, the joint density estimations can generally model the correlation between GW% and Thrust%. In the meantime, we can observe that the current method is still not perfect in dealing with some details, such as the influence from *"Outliers"*.

In the next step, we choose 2 small datasets (December 15, 2014 and February 27, 2015) and propagate uncertainty through AEDT using the estimated densities. Results from the uncertainty propagation are shown in Figure 4.14, 4.15, 4.16, 4.17. The two groups of results are discussed in details in their respective sections.

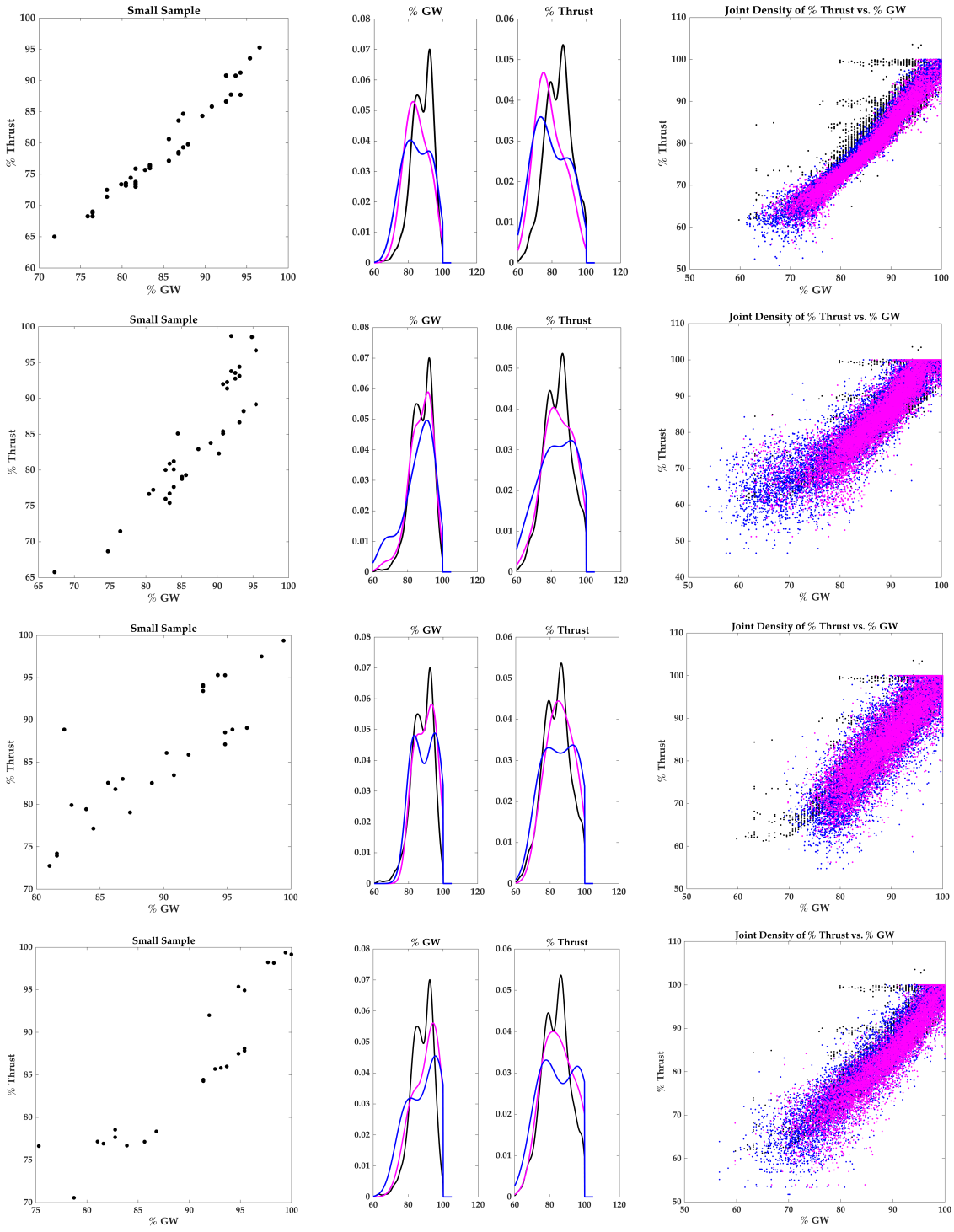


Figure 4.13: Original small samples and their density estimations for September 25, 2014 (first row), June 22, 2014 (second row), December 15, 2014 (third row), and February 27, 2015 (fourth row)

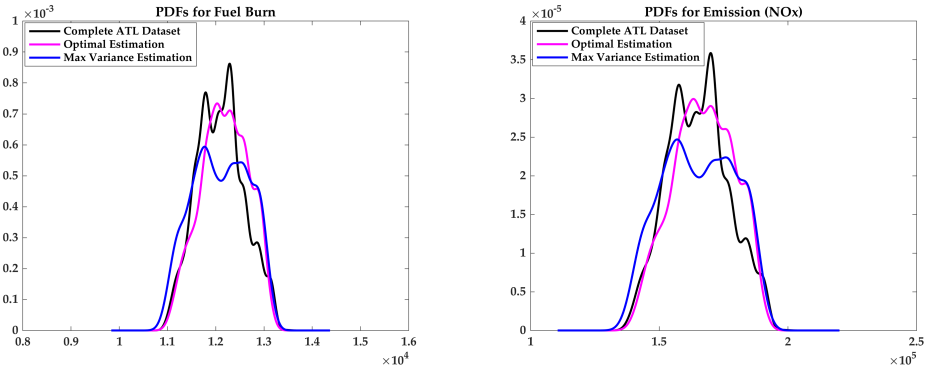


Figure 4.14: AEDT uncertainty propagation result: empirical PDF of fuel burn and emission, using small dataset from December 15, 2014 at ATL (26 flights)

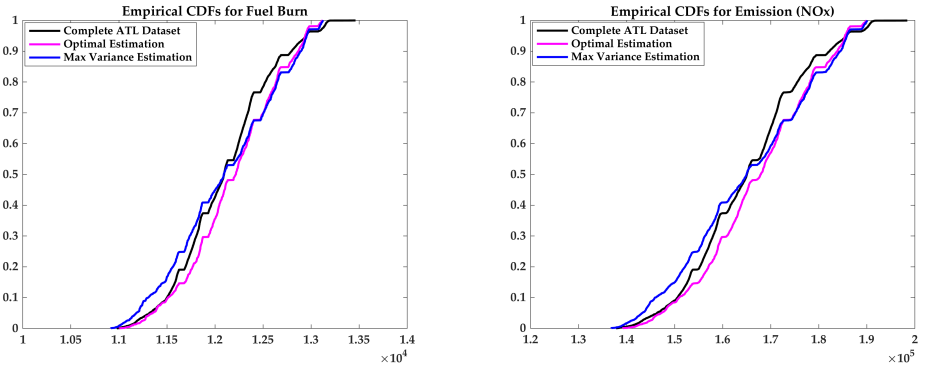


Figure 4.15: AEDT uncertainty propagation result: empirical CDF of fuel burn and emission, using small dataset from December 15, 2014 at ATL (26 flights)

This group of uncertainty propagation results shows some positive sides of estimating one year's fuel burn and emission by using one day's data. Even with as few as 26 flights, the final uncertainty propagation results using the estimated density can give us PDFs that are very close to the true PDFs in statistical moments (Figure 4.16). In the meantime, the empirical CDFs, especially the CDFs of the optimal estimation, look positive (Figure 4.17).

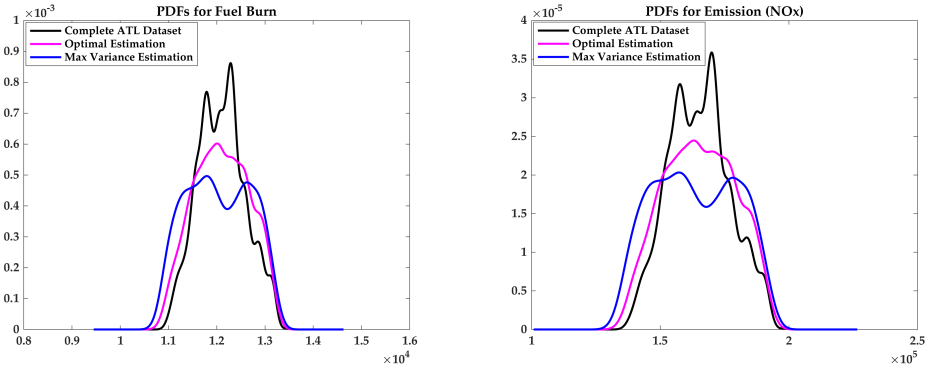


Figure 4.16: AEDT uncertainty propagation result: empirical PDF of fuel burn and emission, using small dataset from February 27, 2015 at ATL (24 flights)

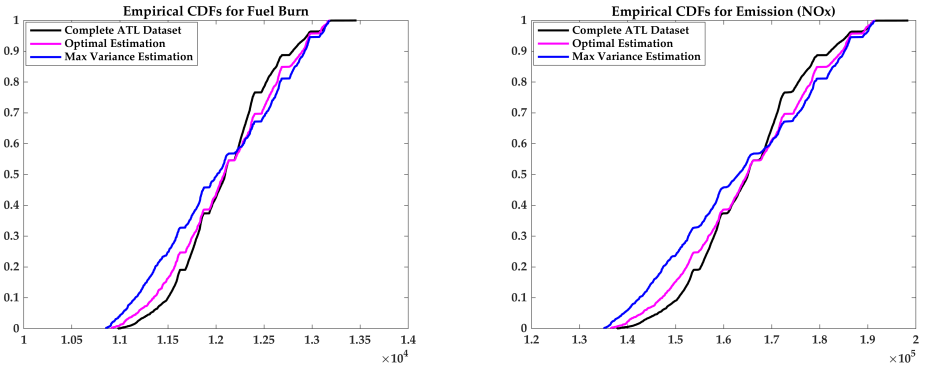


Figure 4.17: AEDT uncertainty propagation result: empirical CDF of fuel burn and emission, using small dataset from February 27, 2015 at ATL (24 flights)

The second group of uncertainty propagations result is not as good as the first group. As can be observed from Figure 4.16, even the optimal PDFs seems to have larger variances than the true distribution, and the shape is also not close. As a result, the optimal CDFs in Figure 4.17 can not provide good enough estimations on the true CDF in the tail regions.

Basically, the two groups of uncertainty propagation results show the two different situations: more positive on the December 15, 2014 group and less positive on the February 27, 2015 group. Nevertheless, since the two days are picked randomly from the whole year, and uncertainty does exists in the small dataset cases, we can say that the uncertainty propagation results are not unexpected.

Effect on Estimating Probability of Success

The necessity of estimating densities based on small datasets can also be illustrated by another case, in the estimation of probability of success (POS). The estimation of probability of success (POS) is an important motivation for getting the output distributions through uncertainty propagation. The left graph of Figure 4.18 contains three different CDF results: using the complete dataset (black), using the optimal densities (blue), and using the small sample (magenta). The right graph of Figure 4.18 contains part of the left graph, in which the left tail region is amplified.

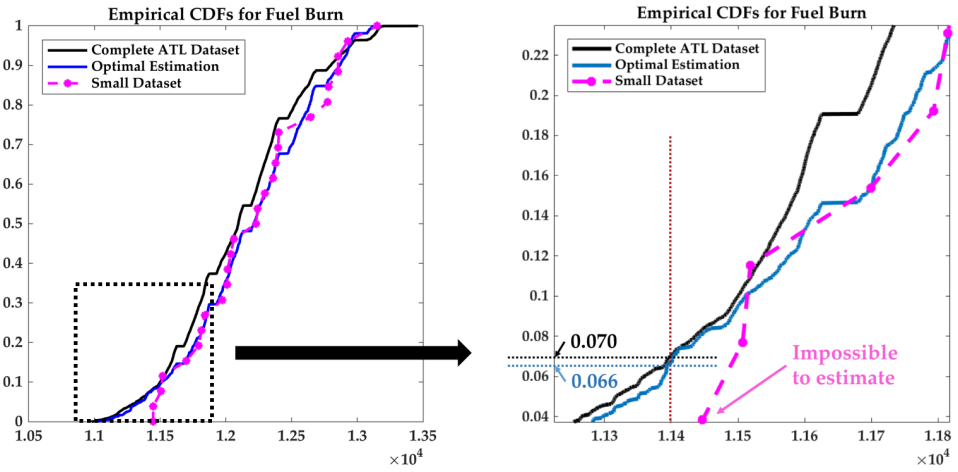


Figure 4.18: Comparison of capability in estimating probability of success from the empirical CDFs by propagation compete dataset, optimal estimation, and the small dataset

When estimating certain POS's from the CDFs, for example, the probability that the fuel burn is smaller than 1.14 $\times 10^4$ kg, the optimal result can provide a valid estimation of 6.6%, which is close to the true POS of 7.0%. As for the small sample

result, due to its nature, however, it is not sufficient enough of cover the tail region. Consequently, this POS can not be estimating by only propagating the small sample without a density estimation.

CHAPTER 5
CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

The results of the test cases demonstrate that the proposed nonparametric-based approach is effective in general in answering the research questions in the first chapter as it can effectively treat uncertainties inherent in small datasets. Below is a re-visit of the overarching research question as well as the four detailed research questions.

Overarching Research Question: *What is an appropriate method to propagate uncertainty of imprecise probabilities due to small dataset while solving the constraints and limitations of the current parametric-based method and staying consistent with our knowledge without introducing unwarranted information?*

Research Question 1: *How can the nonparametric density estimation methods be used to obtain an optimal sampling density to estimate the true density using small dataset and to propagate uncertainty? The key part in the estimation is to identify a nonparametric model without overfitting or underfitting.*

Research Question 2: *Apart from the optimal sampling density, can we provide another more conservative sampling density such that more potential extreme values in the output distributions can be captured through uncertainty propagation process? The necessity for this option exists because uncertainty is an inherent nature of small datasets.*

Research Question 3: *When dependencies among the variables exist, how can the nonparametric-based method be adapted to capture the dependencies and generate multi-dimensional density estimations?*

Research Question 4: *How can the linear correlation between two variables be identified and confirmed based on small dataset?*

All the four research questions have been answered throughout the thesis research, and the overall research objective has also been achieved through the accomplishment of three objectives. The first objective (for Research Question 1) focused on using nonparametric statistical elements to estimate an optimal sampling density for the quantification and propagation of uncertainty created by lack of sufficient statistical data. To obtain such a nonparametric model without overfitting and underfitting, an optimal bandwidth for the Kernel Density Estimation (KDE) is estimated through a combination of classical rules of thumb methods and cross-validation methods. The one-dimensional optimal sampling density estimation method is then tested by four test cases that differ in their characteristics. The test results (both visual results and statistical moments) show that the method is capable of providing good estimations. Besides, the 1-D nonparametric optimal sampling density overcomes constraints and limitations of the parametric-based methods in the following aspects: 1). It does not need people to manually supply a group of candidate probability models based on limited knowledge and judgments; 2). It can estimate a wider range of densities other than just well-established parametric models; 3). The loss of optimality will not happen as more data are added in.

The second objective (for Research Question 2) is to provide another more conservative sampling density other than the optimal sampling density, to represent risk and uncertainty that is inherent in small datasets. By sampling from the more conservative densities for uncertainty propagation, the aim is to simulate more potential extreme values in the output distributions and help making safer decisions. Motivated by the 99% bootstrap confidence band, another sampling density called the maximum variance sampling density is generated through an algorithm that utilizes bootstrap, KDE, a ranking process, and regrouping the data. The four 1-D test cases show that

the maximum variance densities have very good performances in addressing the need, while maintaining close central tendencies with the optimal sampling densities. The complete uncertainty propagation through a systems model experiment shows that output distributions that are obtained through propagating the maximum variance densities have larger variances than the optimal results, and can be used to make safer decisions.

The final objective (for Research Questions 3 & 4) is to model dependencies between variables based on small datasets. After a literature survey on real projects in engineering, linear or near-linear relationship (strong or week) is the main target to model. Aiming to first model 2-D correlation, the task consists of two sequenced parts. The first part is to judge if the linear correlation between two variables really exists, based on the small dataset. Such an identification algorithm is built using various processes, including curvilinear regressions, correlation coefficient, and different types of confidence intervals. After the linear relationship between the two variables is confirmed, the Copulas and Sklar's Theorem is then used to link the marginal densities obtained from the 1-D estimation, to create a joint 2-D estimated density. By sampling from the joint densities for dependent variables, we can prevent uncertainty in the outputs from being underestimated or overestimated. Three different 2-D test cases show that the copula-based 2-D density estimation method is effective in capture the dependency among variables.

Finally, both the 1-D and 2-D estimation methods are applied to a challenging problem in aviation environmental impact analysis. The values of this work is further reflected by the tests results, because the approach developed by this work has the capability such that, if the small datasets/incomplete information is all that we have for uncertainty sources, this approach tries its best to propagate uncertainty with risk consideration based on what we have.

5.2 Contributions

1. Nonparametric density estimates in both 1-D and 2-D that can be used to propagate the uncertainty of small datasets for uncertain inputs, with

    - **No need to manually supply a group of candidate probability models** based on limited knowledge and judgments (No need to make assumptions)
    - **No introduction of any unwarranted information** that is inconsistent with current knowledge
    - **The capability of estimating unconventional densities** (multi-modal, etc.)

2. The introduction of the **maximum variance sampling density** (1-D and 2-D) helps incorporating risk and uncertainty that is inherent with small datasets in the uncertainty propagation process, for making safer decisions

5.3 Future Work

The research conducted in this thesis is a relevant complete work as all the hypotheses tested have been able to fulfill the research questions under certain conditions. Additionally, this topic is designed specifically for the scope and depth of a master's thesis. As a result, the scope of this work limits its possibility of being extended to another master's thesis work or a Ph.D dissertation. Some interesting in-depth explorations lie in the details of statistical modeling, and may require deeper understanding of the theory of statistics. Recommendation for some possible future works are:

- **Selection of the bandwidth for optimal sampling density**: as one of the most challenging parts in nonparametric statistics, bandwidth selection for the KDE has been studied extensively by statisticians in the past, and may

continue to be explored in the future. After a sound literature review, this thesis utilized a combination of rules of thumb methods and cross-validation methods to find an optimal bandwidth. Some second generation methods, such as the solve-the-equation plug-in approach, and the smoothed bootstrap can be further tested. Some researchers had also suggested to use multiple bandwidths in this process. More advanced methods that have been proved to be superior will be a rewarding supplement of the current framework.

- **Density Estimation with Boundary**: in science and engineering research, sometimes the real data are bounded by some logical or physical boundaries. For example, the take-off weight of an aircraft can not exceed the MTOW, and a percentage ratio can not exceed 100%. These boundaries create additional difficulties for density estimation, because from the perspective of model fitting, the estimated density may contain a portion that is outside of the boundaries. To get better uncertainty propagation results, samples of inputs that contain data from impossible region are unaccepted. In the AEDT test case, observing that only very small parts of the estimated densities are actually outside of the logical or physical boundaries, the densities were truncated, and the inside parts were scaled up using conditional probabilities. Methods that can better managing the boundaries in density estimation are beneficial.

- **Multi-fidelity treatments combined with sensitivity analysis**: in order to obtain better estimations for the variables based on small datasets, the calculation of each density estimation involves a large number of runs in parts such as the bootstrap and cross-validation. Even though it only takes a few minutes for each estimation on a PC, if the systems model has numerous input variables (say, $n > 50$) that are needed to be estimated, it is necessary to save running time in this process. In that case, the sensitivity analysis can be used to first

identify the relative influence of each input variable. For those less influential input variables, density estimations with lower fidelities can be applied such that the overall running time can be further saved.

REFERENCES

[1] A. Owen, *Monte Carlo theory, methods and examples: Introduction*. Stanford University Statistics Notes, 2013.

[2] X. Du, *Chapter 8 Monte Carlo Simulation*. Missouri University of Science and Technology, Engineering Uncertainty Repository Notes, 2011.

[3] M. Berger, *Chapter 9 Monte Carlo methods*. New York University Lecture Notes: G22.2112 / G63.2043, Scientific Computing, 2006.

[4] K. G. Schwartz and D. N. Mavris, "Planning technology development experimentation through quantitative uncertainty analysis," *AIAA SciTech Forum - 54th AIAA Aerospace Sciences Meeting, San Diego, CA*, AIAA 2016-0536 2016.

[5] M. J. Asher, B. F. W. Croke, A. J. Jakeman, and L. J. M. Peeters, "A review of surrogate models and their application to groundwater modeling," *Water Resources Research*, vol. 51, pp. 5957–5973, 2015.

[6] MathsNZ, *Part 4: The Triangular Distribution, NCEA Level 3 - 3.14 Probability Distributions*. MathsNZ, 2018.

[7] K. Conrad, "Probability distributions and maximum entropy," *Expository papers, Department of Mathematics, University of Connecticut*, 2018.

[8] J. Zhang and M. Shields, "Efficient propagation of imprecise probabilities," *7th International Workshop on Reliable Engineering Computing (REC2016)*, vol. Ruhr University Bochum, 2016.

[9] F. P. Coolen, M. C. Troaes, and T. Augustin, "Imprecise probability," *International Encyclopedia of Statistical Science*, pp. 645–648, 2011.

[10] T. Zaidi, H. Jimenez, and D. Mavris, "Quantifying random variable dependence structure through copulas theory for probabilistic assessment," *14th AIAA Aviation Technology, Integration, and Operations Conference*, vol. (AIAA 2014-2171), 2014.

[11] J. Zhang and M. D. Shields, "On the quantification and efficient propagation of imprecise probabilities resulting from small datasets," *Mechanical Systems and Signal Processing*, vol. 98, pp. 465–483, 2018.

[12] S. Sankararaman and S. Mahadevan, "Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data," *Reliability Engineering and System Safety*, vol. 96, pp. 814–824, 7 2011.

[13] A. D. Kiureghian, "Analysis of structural reliability under parameter uncertainties," *Probabilistic Engineering Mechanics*, vol. 23, pp. 351–358, 4 2008.

[14] S. Sankararaman and S. Mahadevan, "Distribution type uncertainty due to sparse and imprecise data," *Mechanical Systems and Signal Processing*, vol. 37, pp. 182–198, 1-2 2013.

[15] C. R. Fox and G. Uelkuemen, "Distinguishing two dimensions of uncertainty," *Perspectives on Thinking, Judging, and Decision Making*, 2011.

[16] M. D. Shields, *How much data do I really need to conduct probabilistic UQ?* USACM - Workshop on Uncertainty Quantification and Data-Driven Modeling, 2017.

[17] Y. Xie, *Lecture 12: Monte Carlo Methods and Simulation*. Georgia Tech ISyE 6416: Computational Statistics Spring 2017 lecture notes, 2017.

[18] FAA, *Aviation Environmental Design Tool (AEDT) Version 2c: Service Pack 2 User Guide, March 2017*. Aviation Environmental Design Tool (AEDT) Info, 2017.

[19] S. Geisser and W. Johnson, *Modes of Parametric Statistical Inference*. JOHN WILEY and SONS, INC., 2006.

[20] D. R. Cox, *Principles of Statistical Inference*. Cambridge University Press, 2006.

[21] P. Kvam and B. Vidakovic, *Nonparametric Statistics with Applications in Science and Engineering*. JOHN WILEY and SONS, INC., 2007.

[22] L. Wasserman, *Density Estimation*. Carnegie Mellon University 10/36-702 Statistical Machine Learning Notes, 2017.

[23] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L2 theory," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 57, pp. 453–476, 1981.

[24] MathWorks, *Kernel Distribution*. 2017.

[25] S. J. Sheather, "Density estimation," *Statistical Science*, vol. 19, pp. 588–597, No.4 2004.

[26] A. C. Guidoum, *Kernel Estimator and Bandwidth Selection for Density and its Derivatives*. The kedd Package, version 1.0.3.

[27] B. Silverman, *Density Estimation for Statistics and Data Analysis*. Statistics, Applied Probability, London: Chapman, and Hall, 1986.

[28] G. R. Terrell and D. W. Scott, "Oversmoothed nonparametric density estimates," *Journal of the American Statistical Association*, vol. 80, pp. 209–214, 389 1985.

[29] G. R. Terrell, "The maximal smoothing principle in density estimation," *Journal of the American Statistical Association*, vol. 85, pp. 470–477, 410 1990.

[30] C. R. Loader, "Bandwidth selection: Classical or plug-in?" *The Annals of Statistics*, vol. 27, pp. 415–438, No.2 1999.

[31] Y. Xie, *Lecture 7: Model Selection*. Georgia Tech ISyE 6416: Computational Statistics Spring 2017 lecture notes, 2017.

[32] P. Hall, "Large sample optimality of least squares cross-validation in density estimation," *The Annals of Statistics*, vol. 11, pp. 1156–1174, No.4 1983.

[33] A. W. Bowman, "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, vol. 71, pp. 353–360, No.2 1984.

[34] Y. Xie, *Lecture 10: Bootstrap*. Georgia Tech ISyE 6416: Computational Statistics Spring 2017 lecture notes, 2017.

[35] P. F. Dunn, *Lecture 26: Student's t Distribution (Section 6.4)*. University of Notre Dame AME250 lecture notes, 2018.

[36] R. Griffiths, *Section 6.2: Confidence Intervals for the Mean (Small Samples)*. Mercyhurst University Mathematics lecture notes, 2018.

[37] FAA, *Aviation Environmental Design Tool Version 2b Uncertainty Quantification Report*. Federal Aviation Administration, 2017.

[38] M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research," *Malawi Medical Journal*, vol. 24, pp. 69–71, No.3 2012.

[39] V. Bewick, L. Cheek, and J. Ball, "Statistics review 7: Correlation and regression," *Critical Care*, vol. 7, pp. 451–459, No.6 2003.

[40] eMathZone, *Linear and non linear correlation*, https://www.emathzone.com/tutorials/basic-statistics/linear-and-non-linear-correlation.html [accessed 13 Feb 2018].

[41] J. H. McDonald, *Handbook of Biological Statistics (3rd ed.)* Sparky House Publishing, Baltimore, Maryland, 2014.

[42] J. neng Hwang, S. rong Lay, and A. Lippman, "Nonparametric multivariate density estimation: A comparative study," *IEEE Transactions on Signal Processing*, vol. 42, pp. 2795–2810, 10 1994.

[43] D. W. Scott and S. R. Sain, *Multi-dimensional Density Estimation*. Preprint submitted to Elsevier Science, 2004.

[44] W. Haerdle, M. Mueller, S. Sperlich, and A. Werwatz, *Nonparametric and Semiparametric Models, An Introduction*. Springer, 2004.

[45] T. Duong, *Ks: Kernel density estimation for bivariate data*. cran.at.r-project.org, 2011.

[46] X. Zhang, M. L. King, and R. J. Hyndman, "A bayesian approach to bandwidth selection for multivariate kernel density estimation," *Computational Statistics Data Analysis*, vol. 50, pp. 3009–3031, 11 2006.

[47] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Annals of Statistics*, vol. 38, pp. 2916–2957, 5 2010.

[48] C. Ohlwein, *Multivariate probability distributions: An introduction to the copula approach*. Hans-Ertel-Centre for Weather Research, Meteorological Institute, University of Bonn, Germany, 2014.

[49] P. Bloomfield, *Copulas [Statistics 810-006, Statistics and Financial Risk]*. North Carolina State University Lecture Notes, 2013.

[50] L. Rueschendorf, *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios*. Springer Series in Operations Research and Financial Engineering, 2013.

[51] I. Montes, E. Miranda, R. Pelessoni, and P. Vicig, "Sklar's theorem in an imprecise setting," *Fuzzy Sets and Systems*, vol. 278, pp. 48–66, 2015.

[52] S. Demarta and A. J. McNeil, *The t Copula and Related Copulas*. Department of Mathematics, ETH Zurich, 2004.